Machine Learning Frameworks for Interpreting and Predicting Player Activity Patterns in Gaelic Football

Valerio Antonini

MSc in Data Science

Supervised by Prof Mark Roantree and Dr Michael Scriney



A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

April 2026

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed (Typed Name): Valerio Antonini

Student ID: 21267893

Date: April 21st, 2026

Generative AI Statement

Use of Generative AI in this Thesis

In accordance with DCU's Position Statement on the Use of Artificial Intelligence Tools and the guidelines for responsible use of generative AI in doctoral research, I declare the following use of generative AI tools in the production of this thesis.

Tools

- **Claude (Anthropic)**, Claude Sonnet, used between 2024 and 2026 (claude.ai)

Outputs

The tool was used for the following purposes:

- *Writing Support*: Improving the clarity, grammar, and academic tone of selected sentences and paragraphs across the thesis.
- *Code Development and Debugging*: Identifying and resolving errors in Python code, refactoring for improved readability, and implementing additional functionality to existing scripts.
- *LaTeX Formatting*: Troubleshooting compilation errors and ensuring consistent document structure.

Prompts

- *Writing support*: "Can you improve the quality and clarity of this sentence while maintaining an academic tone" + sentence
- *Code debugging*: "I am getting the following error when running this Python script. Can you identify the issue and suggest a fix" + python script

Iterations

The tool was used on multiple occasions across the duration of the thesis. For writing support, revised sentences were reviewed and accepted, modified, or rejected by the candidate based on accuracy and appropriateness. For code debugging, suggested fixes were tested and verified independently before being incorporated.

Critical Reflection

All research conception, experimental design, data collection, analysis, and scientific conclusions are the candidate's own original work. No research data or unpublished results were shared with the AI system. All AI-generated content was independently reviewed and verified by the candidate before inclusion. The tool served as a supportive assistant for technical and presentational tasks and did not contribute to the intellectual content or scientific findings of this thesis.

Declaration

I declare that while generative AI tools were used to enhance the technical quality, efficiency of production of research and presentation of this thesis, all research conception, experimental design, data collection and analysis, result interpretation, and scientific conclusions are my own original work or the work of my collaborators. The AI tools served primarily as sophisticated assistants for technical improvements to code and writing alongside literature discovery, but did not contribute to the fundamental intellectual content or scientific discoveries presented in this thesis.

Signed (Typed Name): Valerio Antonini

Student ID: 21267893

Date: April 21st, 2026

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Mark Roantree, for his invaluable guidance, unwavering attention, and constructive feedback throughout this intense journey. I am truly grateful for the opportunity he gave me to pursue this PhD, which has shaped both my academic and personal growth.

I extend my sincere thanks to my second supervisor, Dr. Michael Scriney, for his insightful suggestions and engaging discussions, which significantly enhanced the quality of my work.

I am also very thankful to my friend Dermot, who helped me settle into this country and introduced me to GAA and Irish culture. His invaluable support, both within and beyond academia, has meant a great deal to me.

Special thanks go to my friends Sonal, for her constant help over the years, and Aidan, for his invaluable feedback and support at the start of this journey. I would also like to thank all my friends in the research group and at DCU for their kindness and inspiration.

To my Italian friends, both here in Ireland and back home, thank you for making this journey more enjoyable. After long, intense weeks, sharing moments with all of you was the best reward I could ask for.

Most importantly, I want to thank my girlfriend Dee. Your love, incredible patience, and support were fundamental to me, especially while writing this thesis. Your presence was a vital source of happiness and strength.

Finally, I am deeply grateful to my mum Stefania, my dad Bruno, my brother Federico, and the rest of my family for their unwavering love over the years. Despite

the distance, I never felt alone.

List of Publications

1. Antonini Valerio, Alessandra Mileo, and Mark Roantree. "Engineering features from raw sensor data to analyse player movements during competition." Sensors 24.4 (2024): 1308.

2. Antonini Valerio, Dermot Sheridan, and Mark Roantree. "A multidimensional database of in-game player movements (Actions and events) in gaelic football." Data in Brief 57 (2024): 111082.

3. Antonini Valerio, Michael Scriney, Alessandra Mileo, and Mark Roantree. "Using a Spatial Grid Model to Interpret Players Movement in Field Sports." In International Conference on Big Data Analytics and Knowledge Discovery, pp. 395-400. Cham: Springer Nature Switzerland, 2024.

4. Antonini Valerio, Michael Scriney, and Mark Roantree. "Forecasting High-Intensity Activity in Elite Gaelic Football: A Machine Learning Approach Using Wearable Sensor Data and Metabolic Work Ratio". Accepted at the 3rd International Sports Analytics Conference and Exhibition (ISACE), 2026.

5. Sheridan Dermot, Valerio Antonini, Michael Scriney, and Mark Roantree. "Automated Detection of Shot Events in Game Phases Using GNSS Data from a Single Team." In International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 95-105. Cham: Springer Nature Switzerland, 2024.

Contents

Generative AI Statement **ii**

1 Introduction **1**

1.1 Background: From Statistics to Machine Learning in Sports 2

1.2 Gaelic Football: Context and Analytical Challenges 5

1.3 Problem Statement 9

1.4 Hypothesis and Research Questions 12

1.5 Research Contributions 13

1.6 Thesis Structure 14

2 Related Research **17**

2.1 Feature Engineering Sensor Data 17

2.1.1 Velocity Zones and Running Performance Metrics in Team Sports and Gaelic Football 18

2.1.2 From Raw Sensor Data to Structured Features 21

2.1.3 Movement Sequences in Team Sports 23

2.1.4 Summary 26

2.2 Spatio-Temporal Analysis in Team Sports 29

2.2.1 Graph Representations of Player Interactions 29

2.2.2 Spatial and Tactical Analysis Methods 33

2.2.3 Spatio-Temporal Graph Models Beyond Sports 35

2.2.4 Summary 37

2.3 Metabolic Power and Workload Prediction 39

2.3.1 Metabolic Power: Theoretical Foundation 39
2.3.2 Metabolic Power in Gaelic football 42
2.3.3 Machine Learning for Athlete Performance Prediction 43
2.3.4 Summary 47
2.4 Automatic Detection of Game Phases in Sport 48
2.4.1 Summary 52
2.5 Literature Review Summary 53

3 Developing a Threshold-Based Activity Feature Set 56
3.1 Introduction 56
3.2 Raw Sensor Data 58
3.2.1 Data Collection 58
3.2.2 Data Structure and Preprocessing 60
3.2.3 Overall Dataset Characteristics 63
3.3 A Multi-Step Framework for Running Activity Datasets 66
3.3.1 From Wearable Sensor Data to the Action dataset 69
3.3.2 Aggregating Action Data into Event-Level Data 77
3.4 Case Study-Based Validation 83
3.4.1 Case Study 1: Analysis of Velocity Zones using Action Dataset 83
3.4.2 Case Study 2: Analysis of Events and Predicting Distance at High Speed using Event Dataset 92
3.5 Summary 106

4 Interpreting Player Movement Using a Spatial Grid Model 108
4.1 Introduction 108
4.2 Methodology 110
4.2.1 Input Data 111
4.2.2 Spatial Cells Mapping 111
4.2.3 Describing Player Movements as a Multi-Edges Graphs 113
4.2.4 Aggregation into a Simpler Directed Graph 116

4.2.5 Analysis of Field Zones with Graph Metrics . . . 117
4.3 Results . . . 122
4.3.1 Evaluation: Spatio-Temporal Analysis of Game 811 . . . 123
4.3.2 Evaluation: Spatio-Temporal Analysis of Game 893 . . . 129
4.4 Summary . . . 138

5 Forecasting Transitions in Metabolic Workload 140
5.1 Introduction . . . 140
5.1.1 Metabolic Power Framework . . . 141
5.2 Methodology . . . 144
5.2.1 Input Data . . . 146
5.2.2 Feature Engineering . . . 146
5.2.3 Labelling . . . 148
5.2.4 Model Pipeline . . . 155
5.3 Validation and Discussion . . . 161
5.3.1 Experiment 1: Feature Set Evaluation . . . 163
5.3.2 Experiment 2: Model Comparison . . . 164
5.3.3 Insights on Features . . . 167
5.4 Summary . . . 170

6 Detecting Shot Events in Game Phases 172
6.1 Introduction . . . 172
6.1.1 Log Data: Collection and Structure . . . 174
6.1.2 Selecting Shots Phases as Classification Target . . . 175
6.1.3 Experimental Design . . . 177
6.2 Methodology . . . 178
6.2.1 Game Phases Extraction from Log Data . . . 179
6.2.2 Feature Extraction from Sensor Data . . . 181
6.2.3 Model Pipeline . . . 184
6.3 Results . . . 186

6.3.1 Experiment 1: Team A Shots Classification 187
6.3.2 Experiment 2: Team B Shots Classification 189
6.4 Summary 190

7 Conclusion **193**
7.1 Dissertation Overview 193
7.2 Research Questions and Findings 194
7.2.1 Chapter 3: Developing a Threshold-Based Activity Feature Set 194
7.2.2 Chapter 4: Interpreting Player Movement Using a Spatial Grid Model 195
7.2.3 Chapter 5: Forecasting Transitions in Metabolic Workload . . 196
7.2.4 Chapter 6: Detecting Shot Events in Game Phases 197
7.3 Integrated Findings and Practical Implications 198
7.4 Thesis Limitations 199
7.5 Suggestions for Future Research 201

A Terminology **204**

List of Figures

1.1 Overview of the thesis structure. Raw wearable sensor data from elite Gaelic football matches serves as input to a series of data engineering and machine learning frameworks, addressing four research gaps and producing insights and decision support tools for coaches and practitioners. 13

3.1 30 seconds of speed data sampled at 10 Hz for Player 153 during Game 934 62

3.2 30 seconds of tri-axial accelerometer data sampled at 100 Hz for Player 152 during Game 934 63

3.3 Box-plot distribution of speed by half of the game. 65

3.4 Overview of the multi-step framework for transforming raw GNSS and IMU sensor data into the Action and Event datasets. The pipeline progresses from raw sensor data through per-second aggregation, velocity zone labelling, and the DetectAction and DetectEvent algorithms, producing two structured datasets that preserve player movement information. 66

3.5 Sequential event detection visualized through speed data. 79

3.6 Difference in action count between first and second half per game. . . 89

3.7 Probability matrix of action transitions. The starting action is on the y-axis, while the following action is on the x-axis. Each value in the matrix represents the probability of going from the action on the y-axis to the action on the x-axis. 92

3.8 Boxplot of the distribution of event count by half. 95
3.9 **Left**: Frequency of events per duration. **Right**: Frequency of events per duration $\geq$ 30 s. 96
3.10 **Left**: Frequency of events per distance. **Right**: Frequency of events per distance $\geq$ 130 m. 96
3.11 Frequency of events per high-speed distance. **Right**: Frequency of events per high-speed distance $\geq$ 35 m. 97
3.12 The time series of high-speed distance for the remaining games are merged to form a unique time series. The first half of the tested game is placed in the last position. These data constitute the training set for the predictive experiments. 99
3.13 Architecture of the LSTM-based neural network proposed. 101
3.14 Architecture of the CNN-based neural network proposed. 102
3.15 Comparison of true and XGBoost predicted total high-speed distances for each player in Game 788. 105
3.16 Comparison of Player 152's true and XGBoost predicted total high-speed distances across games. 106

4.1 Diagram of steps involved in realizing the Time-Window Spatial Activity Graph. 110
4.2 Coordinates of players' actions in Game 934. 111
4.3 Division of playing area into cells obtained via tessellation algorithm. 112
4.4 Spatial mapping coordinates-to-cells in Game 934. 113
4.5 Example TWG for Game 934, using a simplified network with a few players and spatial cells. 116
4.6 Aggregated graph (TWG') for our running example in Game 934, created by combining the edges from the TWG sample into single weighted edges between cell pairs. 118
4.7 Betweenness centrality values for Game 934 during the first 5 minutes (0–300 seconds). 120

4.8 Communities detected for Game 934 during the first 5 minutes (0–300 seconds). . . . 122

4.9 Spatial grid mapping of the pitch. **Left**: grid points obtained via tessellation algorithm. **Right**: the result of assigning each coordinate to the closest cell. . . . 124

4.10 Betweenness centrality for 5-minute rolling windows: [0, 5) (**top-left**), [1, 6) (**top-right**), [2, 7) (**bottom-left**), and [3, 8) (**bottom-right**). . . . 126

4.11 Community detection for 5-minute rolling windows: [0, 5) (**top-left**), [1, 6) (**top-right**), [2, 7) (**bottom-left**), and [3, 8) (**bottom-right**). Black points represent coordinates not covered by players in the analyzed time window. . . . 128

4.12 Spatial grid mapping of the pitch. **Left**: grid points obtained via tessellation algorithm. **Right**: the result of the association of players' coordinates to their closest grid point. . . . 130

4.13 Spatio-temporal graph depicting detected communities during the first 15 minutes of the game (0, 900 seconds). Node colors represent communities, edge colors indicate average speed, and node sizes correspond to their degree. The label 'goalkeeper' represents the average position of the goalkeeper in the time window. . . . 133

4.14 Spatio-temporal graph depicting detected communities during the time window [900, 1800 seconds), corresponding to the interval from the 15th to the 30th minute of the game. Node colors represent communities, edge colors indicate average speed, and node sizes correspond to their degree. . . . 136

5.1 Methodology diagram illustrating the sequential data processing and analysis steps. . . . 145

5.2 Active and Recovery Areas obtained from metabolic power trend and average metabolic power. . . . 153

5.3 This figure compares each player's total games, games with over 30 minutes played, average minutes per game, and average minutes when participating, highlighting variations in participation and playing time across the team. 163

5.4 This figure presents a comparative analysis of Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), across multiple test games, showing their F1 scores and accuracy to evaluate their predictive performance. 165

5.5 Count of classes in the next time window based on the current class. . 166

5.6 Boxplots showing the distribution of F1 scores per player across games, with colors indicating the number of games played and individual points representing single-game performances. 167

5.7 F1 score of the predictive model across sequential time windows for the first and second half of the matches. 168

5.8 Direction of influence of the features to predict class 1 obtained by averaging shapley values across the tested games. 170

6.1 Example of log data for a game, showing sequential entries with start, click, and end timestamps, along with the relative categories provided by the sports analyst. 176

6.2 Distribution of category frequencies across all games. 176

6.3 Overview of the methodology for shot event detection. Raw GNSS and IMU sensor data and log data are processed through pairwise feature engineering to construct adjacency matrices, which are flattened into feature vectors and fed into machine learning classifiers following RFECV-based feature selection. 179

6.4 Example of extraction and labelling of game phases from log data. . . 180

6.5 Transformation from raw speed data to aggregated pairwise statistics for three players. 183

6.6 Example adjacency matrices representing pairwise metrics between a sample of players during a single game phase. Four of the 16 total matrices are displayed: average Euclidean distance, standard deviation of distance, average speed difference, and average acceleration x difference. Only a small subsample of players is shown. 184

6.7 Confusion matrix for Random Forest classifying shooting events for Team A. 189

6.8 Confusion matrix for XGBoost classifying shooting events for Team B. 190

List of Tables

2.1 Literature review of research in feature extraction from sensor data in sport. 27

2.2 Summary of literature review about spatio-temporal analysis in team sports and beyond. 37

2.3 Comparison of metabolic power assessment and workload prediction approaches in team sports. 48

2.4 Comparison of automatic event and game phase detection approaches in team sports. 53

3.1 Sample of raw wearable sensor data downloaded using Sonra software. Data were recorded at 10 Hz for time, latitude, and longitude, and at 100 Hz for accelerometer (Accl X, Y, Z) and gyroscope (Gyro X, Y, Z) measurements. 62

3.2 Summary of dataset scope and characteristics 64

3.3 Descriptive statistics for player speed (m/s) from the raw data. 65

3.4 Velocity thresholds and corresponding velocity zones used to label player movement intensity, as defined by Rampinini et al. [53] and applied to Gaelic football in this thesis. Each threshold defines the speed range (x_i, in m/s) associated with a discrete movement category, from standing to sprinting. 67

3.5 Sample of the aggregated data. The names of some columns in the header have been shortened: latitude (Lat.) and longitude (Lon.). . . 70

3.6 Sample of the aggregated data with labeled actions. The names of some columns in the header have been shortened: latitude (Lat.), longitude (Lon.), velocity zone (V. Zone), and acceleration (Acc. (m^2/s)). 71

3.7 Features extracted from each action in the Action dataset. For each discrete action summary statistics are computed from GNSS-derived speed and IMU-derived tri-axial acceleration signals, alongside spatial and temporal descriptors. 72

3.8 Sample of the aggregated data after the compression of similar actions. The names of some columns in the header have been shortened: start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), average speed (Avg. Speed (m/s)), and velocity zone (V.Zone). 73

3.9 Direction of movement expressed as labeled turning angles. 77

3.10 Sample of the action data after the compression of similar actions. The names of some columns in the header have been shortened: start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), average speed (Avg. Speed (m/s)), velocity zone (V. Zone), and ID of the direction of movement (Dir. ID). Some of the extracted features are omitted due to space reasons. 77

3.11 Sample of the Action dataset after event detection. The names of some columns in the header have been shortened: player ID (P.ID), start second (Start Sec.), end second (End Sec.), start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), velocity zone (V. Zone), direction (Dir.), duration (Dur.), distance (Dist.), and event ID (E.ID). Some of the columns have been omitted due to space reasons. 81

3.12 Features extracted from each event in the Event dataset. Each event is described by temporal, spatial, speed, acceleration, and movement pattern features derived from the underlying sequence of actions. . . . 82

3.13 Sample of the Event dataset obtained as a final step of the proposed multi-step framework. The names of some columns in the header have been shortened: player ID (P.ID), event ID (E.ID), start second (Start Sec.), end second (End Sec.), avg speed (Avg Sp.), std speed (Std Sp.), duration (Dur.), standard deviation duration (Std Dur.), distance (Dis.), high-speed distance (High-Speed Dist.), and unique turning angle (U.T.A.). Some of the existing columns have been omitted due to space reason. 82

3.14 Average (± standard deviation) distance covered (m/min) by player in each velocity zone in the Action dataset and in Gamble et al. (2019) [54]. 85

3.15 Half by half analysis: action count, average, and maximum duration, sorted by count of actions. 88

3.16 Count, average, maximum duration, and distance covered for each type of action during the sampled games, sorted by average distance. 90

3.17 Summary statistics for each velocity zone. 91

3.18 For each game-half, the table reports the total count of detected events, average and maximum event duration (seconds), average and maximum event distance (meters), and average/maximum distance covered above the high-speed running threshold. 94

3.19 Structure of the transformed dataset used for the prediction of high-speed distance (H.S.D.) covered by players during the second half of a game. The **target feature** to predict is the variable 31^{st} **H.S.D** located at the end of the sequence. 99

3.20 Average Mean Absolute Error (MAE) per model for each tested game. The MAE is computed on the total high-speed distance predicted per player. 104

4.1 Summary of spatial communities with key physical features for the time window 0-15 minutes. For each community, the number of nodes and unique players, average and spread of speed (AvgSpd, SpdIQR) and acceleration (AvgAcc, AccIQR), total high-acceleration events (HighAcc), and proportions of key action intensities: combined Standing/Walking (Stand/Walk), Jogging (Jog), Running (Run), and High-Intensity Running plus Sprint (HIRun_Sprint) are shown. Values represent means or proportions per community. 132

4.2 Summary of spatial communities with key physical features for the time window 15-30 minutes. For each community, the number of nodes and unique players, average and spread of speed (AvgSpd, SpdIQR) and acceleration (AvgAcc, AccIQR), total high-acceleration events (HighAcc), and proportions of key action intensities: combined Standing/Walking (Stand/Walk), Jogging (Jog), Running (Run), and High-Intensity Running plus Sprint (HIRun_Sprint) are shown. Values represent means or proportions per community. 135

5.1 Features extracted from GNSS and IMU variables. 149

5.2 Comparison of predictive performance across different feature sets. . . 164

6.1 Sample data frame illustrating the phases of the game. Columns include *GameID* (ID of the corresponding game), *Start Event* (event initiating the phase), *End Event* (event concluding the phase), *Start Time* (game time in seconds at the beginning of the phase), *End Time* (game time in seconds at the end of the phase), *Team A Shot* (1 if team A shots 0 otherwise), and *Team B Shot* (1 if team B shots 0 otherwise). 181

6.2 Performance metrics of ML models in classifying Team A shooting events. 188

6.3 Performance metrics of ML models in predicting opponent (Team B) shooting events. 190

A glossary of terms and variables used throughout this thesis is provided in the Terminology appendix (page 204).

Machine Learning Frameworks for Interpreting and Predicting Player Activity Patterns in Gaelic Football

Valerio Antonini

Abstract

Wearable sensors in team sports enable analysis of massive amounts of movement data collected during training and competition, with researchers primarily focused on determining physiological demands and running profiles of athletes. Recent advancements in machine learning have expanded these analyses, enabling prediction of physiological metrics and supporting data-driven decision-making by coaches and practitioners. However, research applying machine learning to wearable sensor data remains in early stages. This thesis develops a feature engineering framework that transforms raw sensor data from elite Gaelic football into structured datasets, substantially reducing data size while preserving analytical information and enabling machine learning models to forecast high-speed distance covered by players in the second half of matches. Spatio-temporal graphs of player actions reveal spatial movement patterns across consecutive time windows. To anticipate bursts of player effort, a metabolic power-based metric called the Metabolic Work Ratio is introduced for prediction. Machine learning models predict this ratio in upcoming time windows using features from current and previous periods. Finally, to reduce dependence on video footage and manual annotation, shooting events are classified using only single-team sensor data (without ball tracking or opposing team data) through pairwise movement relationships between players, represented as adjacency matrices of sensor variable differences, which serve as machine learning input features.

Chapter 1

Introduction

Over the past two decades, wearable sensor technology has transformed athlete monitoring in elite team sports, generating millions of data points per match. The integration of machine learning into this data-rich environment has accelerated dramatically, with the global sports analytics market growing from $4.5 billion in 2024 to a projected $14.5 billion by 2030 [1]. Professional organizations across football, basketball, and other team sports have demonstrated measurable improvements in player management and performance optimization using machine learning applications [2], [3].

However, despite this commercial value and the explosion of data and computational capability, a fundamental challenge persists in sports analytics research. The promise of extracting actionable insights from wearable sensor data through machine learning has not been matched by the development of systematic methodologies that enable such work. While researchers have access to unprecedented volumes of movement data captured during competitive matches, critical questions about how to transform these raw sensor streams into meaningful analytical frameworks remain largely unaddressed in the academic literature. The gap between technological capability and methodological foundation has significant implications for the advancement of sports analytics.

This thesis addresses this gap by developing transparent, scientifically rigorous frameworks for four fundamental methodological challenges: systematic transforma-

tion of raw high-frequency sensor data into structured datasets suitable for analysis; spatial analysis methods for understanding where and how movement patterns occur within the playing area; predictive modeling frameworks for forecasting metabolic workload transitions during competitive matches; and automatic extraction of tactical context from movement patterns without manual video annotation. By delivering documented, reproducible frameworks addressing these core challenges, this thesis provides a foundation for advancing sports analytics through open, interpretable methods that facilitate transparency, validation, and actionable insights.

This chapter presents the research context, motivation, and structure for this dissertation.

1.1 Background: From Statistics to Machine Learning in Sports

Sport analytics is defined as the use of statistical methods and predictive models to extract knowledge from historical data to improve players' physical and technical performances and to support sport coaches in making decisions [4]. Data in sports analytics originate from multiple sources: wearable sensors, video analysis, questionnaires, and fitness tests.

In the last twenty years, elite sports have changed significantly, moving from decisions based mainly on intuition and experience to using data-driven approaches to improve performance [5]. This evolution is more than just new technology: it fundamentally changes how sports coaches and practitioners measure and improve athletic ability. Modern sports analytics gained global attention in 2003 with "Moneyball: The Art of Winning an Unfair Game" [6], which described how the Oakland Athletics used statistical analysis to compete with richer teams in Major League Baseball. However, the application of analytical concepts in sports has deeper roots, with researchers in the 1950s and 1980s pioneering the use of mathematical equations and statistical methods to quantify team performance and player contributions [7], [8],

[9]. What guided these early studies was a fundamental principle: systematic analysis of performance data could reveal patterns and opportunities that are difficult to observe and quantify even for experienced coaches.

Between these early statistical foundations and the current era of wearable sensor data, a series of methodological developments in mathematics, statistics, and machine learning fundamentally changed how sports are analyzed and how performance is understood. In the 1990s and 2000s, event-based data collection (tracking discrete actions such as passes, shots, and tackles) enabled the construction of performance indicators that moved analysis beyond simple counts of goals and assists [10]. The application of probabilistic modelling to shot data produced what is now known as expected goals (xG), a metric that assigns each shot a probability between 0 and 1 of resulting in a goal, based on factors such as distance to goal, angle, and contextual variables such as shot type and assist type [11]. The concept emerged independently among several analysts in the early 2010s, with Sam Green at Opta and Brian Macdonald's work presented at the 2012 MIT Sloan Sports Analytics Conference among the earliest documented formulations [11]. From these analytical origins, xG has become one of the most widely adopted metrics in professional football, now used throughout clubs' analytics departments to influence decisions ranging from player transfers to team performance evaluation [11]. It is displayed routinely as a broadcast graphic on television coverage including the BBC's Match of the Day and Sky Sports, and is referenced by managers and pundits as a standard tool for evaluating the quality of chances created and conceded. The trajectory of xG from an analyst's model to a mainstream broadcast statistic illustrates how methodological advances in sports analytics can reshape how the game is described, evaluated, and ultimately played.

A parallel and equally concrete example of how statistical methodology reshaped sport can be found in basketball. The application of expected points per shot analysis led to what is now known as the three-point revolution in the NBA. The expected points per shot, a straightforward probabilistic calculation, demonstrated

that three-point attempts and shots close to the basket yield higher expected value than mid-range two-point shots. Pioneered by Houston Rockets general manager Daryl Morey from 2012 onwards, this data-driven strategy fundamentally changed how the game is played: three-point attempts across the NBA increased from approximately 12.7 per game in 1998 to over 31 per game by the late 2010s, while mid-range shots declined sharply as teams league-wide adopted the same analytical logic [12]. This is one of the clearest documented cases of a statistical insight (expected value) directly altering the tactical structure of a sport at every level.

Today's sports analytics extends beyond these statistical precedents through the introduction of wearable sensor technology during training and games, creating unprecedented availability of high-frequency movement data. Modern Global Navigation Satellite Systems (GNSS) devices combined with Inertial Measurement Units (IMUs) operating at 10Hz and 100Hz, respectively, generate millions of data points per match [13]. Due to their low cost and portability, such sensors have been widely adopted to assess physical activity [14, 15] and external load in team sports [16–18]. These devices record the physical load of athletes during training sessions and games, enabling analysis of match running performance [19] and identification of periods of intense play [20].

At the core of this transformation is the combination of wearable sensor technology, machine learning, and sports science, which together offer new insights into player movement and team dynamics [21]. However, the availability of data does not guarantee improved outcomes: the critical challenge lies in transforming these vast amounts of data into actionable insights that coaches and practitioners can interpret and apply [21, 22]. This is particularly challenging in team sports, where the complexity of coordinated movement, tactical formations, and opponent interactions creates complex analytical demands. In team sports like soccer, rugby, and Gaelic football, athletes engage in continuous, high-intensity intermittent exercise characterized by frequent accelerations, decelerations, changes of direction, and spatial repositioning [19, 23, 24]. With advances in machine learning, the focus has changed

from analyzing past games to also predicting future outcomes, allowing for potential interventions to improve game strategies or player performance [5].

1.2 Gaelic Football: Context and Analytical Challenges

Gaelic Football, one of Ireland's national sports, combines elements of soccer, rugby, and basketball into a unique game that requires exceptional cardiovascular fitness, explosive power, and tactical intelligence. The sport is governed separately for male and female players: the Gaelic Athletic Association (GAA) oversees the male code, while Ladies Gaelic Football, administered by the Ladies Gaelic Football Association (LGFA), follows a largely similar but distinct set of rules and has a separate competitive structure. The physical demand profiles, positional structures, and research literature differ between codes. This thesis focuses exclusively on male Gaelic football. Gaelic football is an amateur sport: players do not receive any payment. The size of the playing field is one of the largest in team sports (130-145 meters long and 80-90 meters wide). The substantial distances and the nature of the game make the sport very demanding for the players. Gaelic football is played at club (sub-elite), college, and inter-county (elite) levels [25]. Matches are composed of two halves of 30 minutes, except for elite inter-county games where the length of each half is 35 minutes. Each team is composed of 15 players, including a goalkeeper, two defensive lines of three players each, two midfielders, and two offensive lines of three players each, with up to five substitutions permitted during a match [26]. Each end of the field has a goal with a crossbar, similar to the one used in rugby. If the ball goes over the crossbar and between the posts, it is worth one point, while if the ball goes under the crossbar and between the uprights, it is a goal and is worth three points [26]. The ball is round, similar to a soccer ball, and may be played using hands or feet, with both short and long passes. Key playing skills include catching high balls, accurate long kicks, passing, solo running by kicking and catching the

ball, and blocking opponents' plays [27].

The body of research on Gaelic football spans both elite inter-county and sub-elite club competition. While this thesis focuses primarily on elite inter-county players, findings from club-level studies are referenced where they offer relevant contextual evidence on the demands and characteristics of the game and are clearly distinguished from inter-county data throughout.

Elite players cover an average of 8,160-9,222 meters per match, with significant differences between playing positions, with midfielders running the greatest distances and performing the most high-intensity running [23], [19]. Analysis of different match periods reveals systematic declines in running performance across quarters, with high-speed running distance decreasing by 8.8%, 15.9%, and 19.8% in the second, third, and fourth quarters compared to the first [19]. The acceleration profile of elite Gaelic football players has been characterized by an average of 166 ± 41 accelerations per match, with players covering approximately 1,563 ± meters at speed of >17 km/h and 524 ± 190 meters at speed ≥22 km/h [24]. The mean acceleration distance averages 267 ± 45 meters, distributed between 12±5 accelerations per 5-minute period [24]. Significant positional differences exist in acceleration demands, with half-backs, midfielders, and half-forwards performing the highest number of acceleration movements, covering greater distances while accelerating compared to full-backs and full-forwards [28]. These findings establish that Gaelic football creates position-specific physical demands that change during the match [19], [24], [28].

To contextualize these running and acceleration demands within the broader landscape of field-based invasion sports, elite rugby union players (professional club and international level) cover approximately 6,715–6,953 meters per match [29], [30], with backs spending approximately 5.6–7.5% of total match distance in high-intensity running zones (18–24.1 km/h) [30], and performing approximately 85–89 acceleration efforts exceeding 1.5 m/s^2 per match [29]. These values are notably lower than those recorded in elite inter-county Gaelic football. In comparison, Aus-

tralian Football League (AFL) players cover substantially greater distances of approximately 11,000–13,000 metres per game, with midfield players demonstrating greater physical demands than forwards and defenders [31, 32]. Approximately 65% of total AFL match time is spent at low intensities below 8 km·h^{-1}, while players perform roughly 240 moderate accelerations (>4 km·h^{-1} in 1 s) and spend approximately 5 minutes running above 18 km·h^{-1} per game [31]. Match intensity in AFL has evolved across discrete periods, with rule changes, interchange rotations, and evolving coaching strategies identified as primary drivers of fluctuations in both volume and intensity [32], further highlighting how the physical demands of field-based invasion sports are shaped by sport-specific contextual factors.

Research examining performance indicators at the senior club (sub-elite) level has revealed that six variables affect significantly the difference between winning and losing teams, with three directly related to scoring (points, number of scores, and total score) and three related to the effective use of possession (possession/scores ratio, turnover rate, and scores per possession) [33]. Post-match recovery studies have demonstrated that competitive Gaelic football matches at the senior club level (sub-elite) cause significant neuromuscular and perceptual disturbances, with perceptual responses declining by 27.6% and countermovement jump height decreasing by 3.9% immediately at full-time. The full recovery of neuromuscular variables occurred at 48 hours post-match despite prolonged elevation in creatine kinase levels [34]. A systematic review and meta-analysis of injury epidemiology found that the total injury incidence in Gaelic football was 10.7 injuries per 1000 hours of exposure, with match incidence substantially higher at 55.9 injuries per 1000 hours compared to training at 4.6 injuries per 1000 hours, and the lower limb accounting for over 70% of all injuries, with hamstring injuries representing 22-24% of total injuries [25].

While earlier studies have described the key physical and tactical demands of Gaelic football, recent research methods focus on more advanced analysis of player and team performance. This methodological advancement reflects a trend in sports science, where traditional statistics are being complemented by data mining tech-

niques and machine learning or deep learning applications. In Gaelic football, researchers are increasingly applying advanced data science and sports analytics methods to obtain deeper insights from athletes' data, moving from basic descriptions to approaches that can explain and predict performance.

A modified Elo rating system adapted for inter-county competition demonstrates how established statistical frameworks from other domains can be successfully applied to Gaelic football contexts, in this case correctly forecasting match outcomes in 72.9% of cases across a six-year period [35]. The original Elo rating system assigns chess player ratings that update after each match proportional to the difference between actual and expected results [36]. Unsupervised clustering techniques, specifically spectral clustering, combined with the elbow method, have been used to derive Gaelic football-specific velocity thresholds for referees, replacing generic categorizations with data-driven thresholds tailored to the unique demands of the game [37]. The logistic regression model has emerged as a valuable technique for quantifying shot quality in Gaelic football, with features incorporating spatial variables (distance, angle, field location), contextual factors (pressure, shot type), and individual characteristics to predict scoring probability [38], [39]. In this research, *features* refer to computed attributes derived from raw sensor data using mathematical operations (such as summation, averaging, or finding maximum values). Features can represent properties of time segments (for example, duration in seconds, distance covered in meters, maximum speed in m/s) and typically correspond to columns in a dataset. Throughout this thesis, the terms "features" and "metrics" are used interchangeably to describe these attributes.

Machine learning approaches were used to model the relationship between external training load metrics and perceived exertion, with random forest showing the best results for predicting players' perceived load responses using GNSS-based and wellness variables [40]. Most recently, advanced machine learning techniques have been applied to address the problem of limited sample sizes in elite athlete monitoring. Synthetic data generation using Tabular Variational Autoencoders has

been used to augment real-world datasets, enabling more robust model training for performance attenuation prediction [41]. Complementary work has explored the application of data balancing techniques (SMOTE, ROSE, ADASYN) and explainable AI methods (SHAP values, principal component analysis) to identify key predictors of post-match performance decline from multimodal athlete monitoring data [42].

Despite the number of studies in the last years, Gaelic Football remains underexplored compared to other invasion field-based team sports [43], like soccer, rugby, or Australian football. Due to the unique characteristics of the game, some of the analytical frameworks developed for other sports may not work directly in this context. Furthermore, most Gaelic Football analytics research has focused on descriptive statistics and summary performance metrics, with limited exploration of data mining methods and machine learning applications. This gap represents a challenge: developing analytical methods specifically designed for Gaelic Football's distinctive movement patterns, tactical structures, and data availability limitations.

1.3 Problem Statement

The adoption of wearable sensor technology in Gaelic football has grown substantially in recent years. By 2020, the majority of elite inter-county male GAA teams were wearing athlete monitoring devices [44], with Ladies Gaelic football experiencing a 235% increase in GPS technology usage between 2021 and 2024, compared to a 137% increase in the men's game [45]. This widespread adoption has generated unprecedented volumes of high-frequency movement data across the sport. Despite the wide adoption of wearable sensor technology in elite team sports, the current analytical approaches suffer from four fundamental limitations that prevent coaches and sport scientists from fully leveraging sensor data for performance optimization.

1. **The Methodological Transparency Problem**. While wearable sensors combining GNSS and IMU technologies are widely used in team sports, generating millions of data points per match, academic research and commer-

cial applications lack transparent, reproducible methodologies for transforming continuous sensor data into structured datasets suitable for analysis. In this research, a structured dataset refers to a tabular data representation where each observation corresponds to a movement segment defined by specific criteria (for example, a continuous movement within a velocity zone) rather than individual sensor readings. In this context, structured datasets are characterized by dimensional reduction from high-frequency sensor data into discrete units and summary statistics and features computed for each unit.

Existing studies focus on which features to extract (velocity zones, accelerations, distance metrics) but fail to provide the methods to organize raw data to support diverse analytical objectives [23], [24], [46]. This causes significant issues for reproducibility in sports analytics. Researchers may extract summary metrics produced by the sensor companies' software or work directly on the raw data. The formula to compute these metrics is known, but how the software organizes the raw sensor data to produce them is not clear. These summary values offer practical insights, but they only provide a broad overview of a training session or game, limiting detailed analysis beyond aggregate values. When using raw sensor data, researchers usually extract features, but there is no step-by-step framework in the literature that clearly explains how to organize and analyze these large, complex data streams. The raw sensor data can be extremely large, sometimes containing millions of rows and multiple dimensions, which makes managing and processing it even more challenging.

2. **The Spatio-Temporal Movement Problem**. Existing performance analysis in Gaelic football and similar invasion-based team sports focuses almost exclusively on describing when and how intensely players move, but it lacks analysis of where these movements occur within the playing area and how spatial usage patterns evolve during matches. This lack in the spatio-temporal analysis of player movements in the field prevents answering critical questions: which zones of the playing area serve as key transition corridors? How

do regions of coordinated team activity form and dissolve during matches? Do different spatial zones exhibit distinct workload characteristics? Existing graph-based approaches in sports mostly model player-to-player interactions through passing networks, treating players as nodes, capturing tactical relationships, but are unable to reveal how teams collectively move through and use the space.

3. **The Predictive Analytics Problem**. Current sports analytics in Gaelic football mostly focus on analyzing past events, describing what happened rather than anticipating what will happen next. While metabolic power frameworks have validated ecological validity for quantifying energetic demands [47, 48], and researchers have begun exploring machine learning for performance prediction [40, 42], these approaches exhibit a critical limitation: they predict aggregate match outcomes (will a player underperform overall? What is post-match fatigue status?) or next-day wellness rather than imminent metabolic workload transitions that could help coaches to make data-driven decisions during matches, assuming availability of real-time raw data from wearable monitoring systems.

4. **The Tactical Context Gap**. Wearable sensor data in sports often lacks tactical context, because collecting this data requires costly, time-consuming manual video analysis. Analysts must re-watch entire matches and manually code key moments like possession changes and scoring chances, which demands specialized software, significant time, and financial resources. This creates an important limitation: the ability to interpret sensor data in tactical contexts is limited to organizations with dedicated video analysis teams. As a result, most sensor data is used only for physical load management. Current automatic event detection relies on expensive resources like ball tracking or full video footage, which most clubs cannot afford. Without these, movement data cannot be linked to tactical context, limiting understanding of the reasons behind physical demands.

As a result of these four problems, sports organizations possess unprecedented data on athletes but lack reproducible methodologies to process sensor data systematically, frameworks to reveal spatial-tactical organization, capability to anticipate imminent performance changes, and methods to automatically connect physical demands to game situations.

1.4 Hypothesis and Research Questions

The central hypothesis on which this research dissertation is based can now be articulated as follows:

High-frequency wearable sensor data (GNSS and IMU) from elite team sports can be systematically transformed into performance insights through data engineering and machine learning techniques, addressing critical gaps in methodological transparency, spatio-temporal movement understanding, predictive capability, and automatic context extraction, thereby enabling scientists and coaches to make data-driven decisions for performance optimization.

This dissertation addresses four research questions corresponding to the stated hypothesis:

1. **RQ1**: Can raw high-frequency GNSS and IMU sensor data be systematically transformed using documented, reproducible algorithms into structured datasets that preserve movement information and enable both movement profiling and predictive modeling?

2. **RQ2**: Can graph-based representations of player movements through spatial cells reveal dynamic spatio-temporal movement patterns, including critical transition zones, spatial communities with distinct workload profiles, and temporal evolution of these patterns across match phases?

3. **RQ3**: Can machine learning models predict player-specific transitions to elevated metabolic workload in future temporal windows, using novel metabolic

power metrics and activity features from the current window and cumulative game data?

4. **RQ4**: Can shooting events for both teams be automatically classified using only single-team wearable sensor data and pairwise player movement features, eliminating the dependence on manual video-based event annotation?

These four research questions reflect different aspects of the broader challenge of extracting meaningful insights from high-frequency sensor data. The first research questions address systematic data transformation and spatio-temporal analysis using structured datasets derived from raw sensor data (RQ1→RQ2), while the latter two focus on predictive modelling and automatic event detection working directly with raw data (RQ3, RQ4).

Figure 1.1 summarizes the overall thesis approach: wearable sensor data is collected and processed using data engineering and machine learning methods, targeting four identified research gaps, to produce insights and decision support tools for coaches and practitioners.

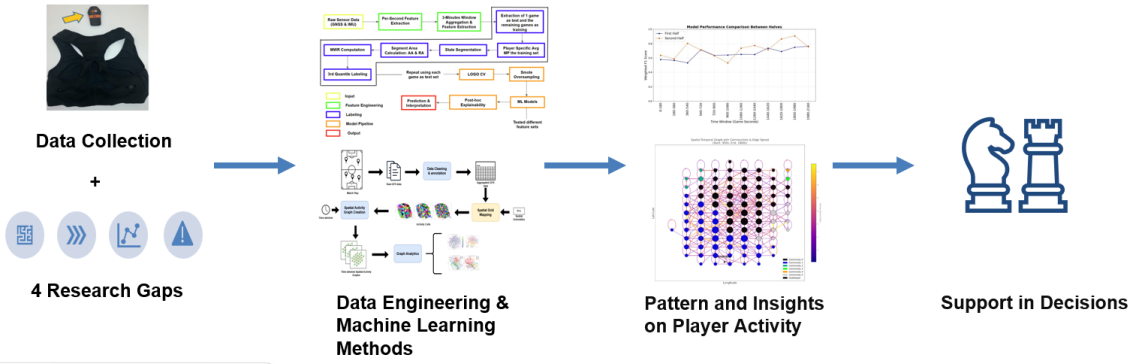


Figure 1.1: Overview of the thesis structure. Raw wearable sensor data from elite Gaelic football matches serves as input to a series of data engineering and machine learning frameworks, addressing four research gaps and producing insights and decision support tools for coaches and practitioners.

1.5 Research Contributions

This thesis makes the following contributions to sports analytics and machine learning research:

Systematic Feature Engineering Framework: A complete, documented methodology for transforming raw GNSS/IMU sensor data into structured Action and Event datasets. The framework addresses the methodological transparency problem by providing reproducible algorithms that enable dimensionality reduction from 76+ million raw observations per match to interpretable movement units while preserving analytical information.

Spatial-Temporal Graph Analytics: A novel cell-based graph framework where spatial cells (rather than players) serve as nodes, enabling identification of critical transition zones through betweenness centrality, detection of spatial communities with quantified workload profiles (speed, acceleration, intensity distributions), and tracking of temporal evolution across match phases. This methodological innovation addresses the fundamental limitation of player-centric graph approaches that cannot reveal collective spatial organization patterns.

Predictive Metabolic Workload Modeling: A machine learning framework forecasting player-specific transitions to elevated metabolic demand states within 3-minute temporal windows (accuracy 0.71 $\pm$0.04, and F1 score 0.69 $\pm$ 0.05).

Sensor-Based Event Classification: An automatic classification system detecting shots made by both teams using only single-team wearable sensor data (accuracy 0.63-0.69, F1 score 0.69-0.68), demonstrating that tactical context can be extracted from movement coordination without ball tracking or video analysis. Here, pairwise player movement relationships are represented as adjacency matrices that serve as input features for machine learning models.

These contributions advance both sports analytics methodology and the broader challenge of extracting meaningful insights from high-frequency spatio-temporal data.

1.6 Thesis Structure

This chapter presented the background of the study, described the current limitations, and highlighted the hypothesis along with the research questions that moti-

vated the research.

Chapter 2 provides a comprehensive review of related work across four research domains: feature engineering approaches for sensor data transformation, graph-based and spatial-temporal analysis methods in team sports, metabolic power frameworks and running performance prediction, and automatic playing event detection. The chapter identifies specific gaps in each domain that motivate the research questions.

Chapter 3 presents the systematic methodology for transforming raw GNSS and IMU sensor data into Action and Event datasets. The chapter provides complete algorithmic specifications, validation against established Gaelic football running profiles, case studies profiling actions and events, and predictive modeling for high-speed running distance. Both the framework and the resulting Action dataset have been published in [49] and [50]. The main contribution is a transparent, reproducible framework with documented algorithms that addresses the methodological transparency problem, enabling dimensionality reduction from 76+ million raw observations to interpretable movement units while preserving analytical information.

Chapter 4 introduces the Time-Window Spatial Activity Graph approach for analyzing spatial-temporal movement patterns. The chapter describes graph construction methodology, applies betweenness centrality and community detection algorithms, and presents case studies revealing tactical patterns and workload distributions across playing areas. This study was presented in [51]. The main contribution is a novel graph-based method using spatial cells as nodes, enabling the identification of transition zones, spatial communities with distinct workload profiles, and their temporal evolution over time.

Chapter 5 develops the predictive framework for forecasting transitions in player metabolic demands. The chapter introduces the Metabolic Work Ratio metric, describes feature engineering and model development procedures, validates prediction performance, and provides explainable AI analysis of key predictive factors. This work has been submitted for publication. The main contribution is a machine learn-

ing system predicting imminent (3-minute window) metabolic workload transitions with accuracy 0.71 $\pm$0.04 and F1 score 0.69 $\pm$ 0.05, enabling real-time coaching interventions (assuming availability of data in real time) rather than retrospective analysis.

Chapter 6 investigates the extraction of shot phases from sensor data alone. The chapter describes the feature engineering approach that constructs adjacency matrices from pairwise player movement differences (distance, velocity, acceleration, gyroscope) and uses these matrices as input features for machine learning models, presents classification results for both own-team and opponent shooting events, and discusses implications for reducing dependence on manual video analysis. This study was presented in [52]. The main contribution is an automatic event classification method using only single-team sensor data (achieving accuracy 0.63-0.69 and F1 score 0.69-0.68), demonstrating that tactical context can be extracted from movement coordination patterns without ball tracking, opponent data, or manual video analysis.

Chapter 7 synthesizes contributions made in each research section and identifies promising directions for future research in sports analytics and machine learning applications for team sports.

Chapter 2

Related Research

This chapter reviews research literature relevant to the four primary research topics addressed in this thesis. Section 2.1 examines feature engineering approaches for transforming raw sensor data into features describing player activity. Section 2.2 reviews graph-based and spatial-temporal analysis methods in team sports and other domains. Section 2.3 discusses the metabolic power framework and machine learning approaches for running performance prediction. Section 2.4 explores automatic playing event detection using machine learning techniques.

2.1 Feature Engineering Sensor Data

This section examines how researchers transform raw wearable sensor data into analytical features, revealing a critical gap: despite widespread wearable sensor adoption, the literature lacks systematic, reproducible methodologies for structuring continuous sensor data into datasets that support multiple analytical objectives.

The transformation of raw wearable sensor data into meaningful features represents a foundational challenge in sports analytics. Torres et al. [21] provide a comprehensive narrative review establishing that tracking systems in team sports serve three primary purposes: describing locomotor characteristics and external load, assisting in the programming of team training external load, and supporting decisions about performance and injury risk. The authors emphasise a critical gap between

technology deployment and analytical practice. Wearable sensors collect continuous athlete external load data, and practitioners use derived metrics to describe, plan, monitor, and evaluate training and competition characteristics. Selecting appropriate metrics depends on the sport's context and demands evaluation by practitioners [21]. Moreover, analyzing the physical output of team sport athletes using overall summary metrics of the entire training or game (such as, for example, distance covered or distance covered at a certain speed) presents challenges, as important periods of changing intensity can be missed. Methods such as moving average intervals or time series segmentation have been developed to detect these changes over variable time segments without fixed intervals [21].

2.1.1 Velocity Zones and Running Performance Metrics in Team Sports and Gaelic Football

The segmentation of movements based on velocity thresholds into discrete zones has become a standard analytical approach in team sports performance analysis. This discretization approach enables comparison of running performance across players, positions, and competitions. However, the specific threshold values defining velocity zones vary substantially across sports and research contexts, reflecting differences in sport-specific demands, player populations, and methodological conventions.

Early foundational work by Rampinini et al. [53] proposed velocity thresholds for general team sports that have subsequently been adopted in Gaelic football research [54]: standing ($\leq$0.194 m/s, $\sim$0.7 km/h), walking (0.194–2 m/s, $\sim$0.7–7.2 km/h), jogging (2–4 m/s, $\sim$7.2–14.4 km/h), running (4–5.5 m/s, $\sim$14.4–19.8 km/h), high-intensity running (5.5–7 m/s, $\sim$19.8–25.2 km/h), and sprinting (>7 m/s, >25.2 km/h). These velocity zones categorize speeds values and have become widely adopted across team sports [54] and [55]. However, a universally accepted set of velocity thresholds has not been established across Gaelic football studies, as evidenced by the differing velocity zones applied across the studies reviewed here [23], [24], [54].

Malone et al. [23] studied the positional match running performance of elite male Gaelic footballers using 4Hz GNSS data collected from 50 players across 30 competitive games (215 full game datasets). They classified running activity by total distance, high-speed running distance ($\geq$4.72 m/s or 17 km/h), sprint distance ($\geq$6.11 m/s or 22 km/h), mean and peak velocities, and number of accelerations. The average total distance covered per match was 8,160 $\pm$ 1,482 m, with 1,731 $\pm$ 659 m covered at high speed and 445 $\pm$ 169 m covered in sprints (approximately 44 sprint actions). The peak velocity recorded was 8.42 $\pm$ 0.50 m/s (30.3 $\pm$ 1.8 km/h), and the mean velocity was 1.81 $\pm$ 0.33 m/s (6.5 $\pm$ 1.2 km/h). Players executed 184 $\pm$ 40 accelerations, equating to around 2.6 accelerations per minute. Significant positional differences were found, with midfielders covering greater total and high-speed running distances compared to backs and forwards ($p < 0.001$). A decline in high-speed and sprint distances was observed from the first to the second half, with midfielders, half-backs, and half-forwards experiencing the most pronounced decreases. This foundational study highlights the impact of playing position on running demands in Gaelic football and is one of the first comprehensive descriptions of match running performance in the sport.

In follow up research, Malone et al. [19] examined the running performance and its temporal decrement across quarters in elite Gaelic football using 4Hz GNSS data from 3 inter-county teams over 3 seasons, collecting 250 full-match datasets. Running metrics included total distance (TD), high-speed running distance ($\geq$4.72 m/s or 17 km/h), sprint distance ($\geq$6.11 m/s or 22 km/h), accelerations, and peak speed. The mean total distance covered was 8,889 $\pm$ 1,448 m, with high-speed running averaging 1,596 $\pm$ 594 m. They observed a consistent reduction in TD of 4.1%, 5.9%, and 3.8% in the second, third, and fourth quarters respectively, accompanied by greater decrements in high-speed running distance of 8.8%, 15.9%, and 19.8% compared to the first quarter ($p < 0.001$). Positional analysis indicated that half-backs, midfielders, and half-forwards covered the greatest distances and displayed the largest decrements in running performance across quarters. The peak

velocity recorded was 8.42 ± 0.50 m/s (30.3 ± 1.8 km/h). These findings confirm that exercise intensity declines over time and changes by position.

Ryan et al. [24] quantified the acceleration profile of elite Gaelic footballers using 4Hz GNSS data collected from 36 male players. Acceleration events were defined as instances where a player changed speed by 0.56 m/s (2 km/h) within 1 second, for a minimum of 2 seconds. Acceleration was considered finished once the player decelerated to less than 75% of the maximum speed reached during the prior acceleration effort. Players performed an average of 166 ± 41 accelerations per match and covered 1,563 ± 605 m distances at high-speed running (≥4.72 m/s or 17 km/h] and 524 ± 190 m at very high-speed running (≥6.11 m/s or 22 km/h). Positional analysis indicated midfielders performed significantly more high-intensity accelerations than other positions ($p \leq 0.05$). The study also documented declines in acceleration distances ranging from 2% to 32% throughout the game depending on position. These findings highlight the critical role of acceleration metrics in capturing the high-intensity movements distinctive to Gaelic football.

Cullen et al. [56] investigated the physiological profile and activity patterns of minor (U-18) Gaelic football players. They used velocity zones adapted from previous literature [29], defining walking as speeds below 2.0 m/s (0–1.67 km/h), jogging as speed in 1.67–3.33 m/s (6-12 km/h), cruise running as 3.33–3.89 m/s (12-14 km/h), striding as 3.89–5.00 m/s (14-18 km/h), high-intensity running as 5.00–5.56 m/s (18-20 km/h), and sprinting as speed ≥5.56 m/s (20 km/h). Players were considered to have entered a velocity zone after maintaining the relevant velocity for at least 2 seconds.

Given this heterogeneity in threshold definitions, this research adopted the comprehensive velocity zone framework proposed in [53] and subsequently validated in Gaelic football [54]. Unlike many Gaelic football studies that focus only on high-speed thresholds, authors in [53] provide comprehensive categorization across the full range of movement intensities, from standing to sprinting. This choice enables segmentation across the entire activity profile and provides consistency with prior

Gaelic football research. Furthermore, the same velocity thresholds were adopted for sports similar to Gaelic football, such as Australian Football [55].

2.1.2 From Raw Sensor Data to Structured Features

The transformation of raw GNSS time-series data into meaningful features represents a critical methodological challenge in sports analytics. While high-frequency GNSS devices generate extensive positional and velocity information, converting these streams into structured, interpretable features requires systematic approaches. This subsection reviews related studies addressing feature extraction, data aggregation, statistical summarization, and machine learning applications for transforming raw GNSS data into features for sports analytics.

Imbach et al. [57] investigated the use of raw GNSS data to model athletic performance in elite football players, focusing on predicting individual acceleration-velocity profiles. The acceleration-velocity profile is a function describing how a player's acceleration capacity decreases as their velocity increases during sprinting. It is obtained by fitting a linear regression between acceleration and velocity over a range of speeds (velocities above 3 m/s). Their methodology involves extracting a large set of features from raw GNSS time series using the `tsfresh` Python module [58], capturing statistical metrics such as mean, standard deviation, entropy, and frequency domain characteristics across both time and frequency domains. These features represented the input data to the multivariate machine learning models employed, including ridge regression and long short-term memory networks. The extracted time-series statistical features, while comprehensive, may be difficult for sport coaches to interpret directly, potentially limiting practical application. In contrast, the present thesis develops a versatile, multi-step feature extraction pipeline that integrates both GNSS and IMU data into discrete action datasets specifically tailored to analyze running demands of Gaelic football players.

Midoglu et al. [59] presented *SoccerMon*, the largest publicly available dataset integrating both subjective and objective health and performance metrics collected

from two elite women's soccer teams over two competitive seasons. The dataset encompasses over 33,800 subjective reports covering player wellness, training load perception, injuries, illnesses, and match performance, with more than 10,000 records derived from wearable GPS tracking devices capturing positional, acceleration, rotational, and heart rate data. The *SoccerMon* represents a dataset combining subjective wellness with movement metrics, but does not provide a reproducible pipeline to convert sensor data into a structured dataset.

Hussain et al. [60] proposed a comprehensive feature engineering approach for recognizing and classifying gym exercises using wearable sensor time series data. The study uses IMU data collected from wrist-worn devices during various gym-based activities, aiming to accurately distinguish between exercise types using feature engineering and machine learning techniques. The feature extraction process involves segmenting the raw time series data into windows aligned with possible exercise repetitions and computing a rich set of statistical and domain-specific features from each segment. These features include simple statistical metrics (mean, max, min, standard deviation), frequency-domain attributes (spectral energy, dominant frequency), and signal structure indicators (zero crossings, variance, skewness). The resulting feature vectors are then used to train supervised classifiers that achieve high recognition accuracy for exercise type prediction. Our research builds on similar feature engineering principles but addresses a different problem and domain-specific application: segmenting continuous GNSS and IMU time series from elite Gaelic football into datasets describing actions and events. Our segmentation uses sport-specific velocity thresholds to define meaningful movement categories and aggregates consecutive velocity zones, with statistical features describing those segments. Thus, while the statistical techniques for processing sensor data are similar, the problem addressed in this thesis requires a different segmentation logic.

Fister et al. [61] introduced a minimalistic toolbox for efficient feature extraction from raw sport activity data captured by wearable sensors. The toolbox supports extraction of classical statistical features (mean, median, variance, skewness), tem-

poral features through fixed sliding window segmentation, and frequency-domain features. Its design emphasizes scalability, computational simplicity, and broad applicability across various sports data formats. In contrast, the framework proposed in Chapter 3 provides a detailed step-by-step approach that transforms continuous high-frequency GNSS and IMU sensor data into structured, domain-specific datasets representing discrete player movements in Gaelic football.

Chapter 3 extends these feature engineering approaches by contributing both a transparent algorithmic methodology and resulting structured datasets that integrate GNSS and IMU data, addressing the literature's lack of systematic frameworks for transforming raw sensor data into datasets for Gaelic football analysis.

2.1.3 Movement Sequences in Team Sports

The analysis of discrete events and characteristic movement sequences in team sports represents a complementary analytical approach to traditional aggregate time-window analysis.

Sweeting et al. [62] proposed that characteristic recurring movement sequences could be automatically extracted from spatio-temporal tracking data. The authors calculated velocity, acceleration, and angular velocity from positional (X, Y) coordinates collected at 10 Hz, then grouped each variable's values using one-dimensional k-means clustering. Each cluster was assigned a unique alphabetic label, creating a symbolic representation where continuous movement variables were transformed into discrete movement units. Combinations of these labelled velocity, acceleration, and angular velocity movements were then compared using the Levenshtein distance, with similarities computed by the longest common substring problem to identify frequently recurring sequences. Levenshtein distance is a metric in computer science that measures the smallest number of single-symbol edits (insertions, deletions, and substitutions) needed to change one string into another [63]. The authors demonstrated that closed contiguous movement patterns more effectively separated players into positions (91.02% classification accuracy) compared to non-

consecutive patterns, establishing that movement continuity contains positional information. However, their approach focused on pattern mining rather than comprehensive dataset structuring, treating movement sequences primarily as patterns to extract rather than as features for broader analysis. In contrast, our study complements similar pattern extraction by providing structured datasets that preserve temporal sequences and enable analytical tasks.

White et al. [64] developed the Sequential Movement Pattern-mining framework to extract movement patterns from GNSS data in team sports. After calculating turning angles using bearing values between consecutive coordinates, speed, acceleration, and turning angle are divided into qualitative categories (walk, jog, run, sprint for speed; deceleration, neutral, acceleration for acceleration; and straight, acute-change, large-change, backwards for turning angles) based on thresholds and assigned unique alphabetic labels to each category. They isolated movements into subsequences every time a speed value was lower than 1.20 m/s. These labeled movement units are then aligned chronologically to create sequential movement patterns. The framework includes a condensing step that removes consecutive repeated movement units, leaving only unique transitions in each sequence. By employing the Levenshtein distance to compare sequences and longest common subsequence algorithms to identify frequent patterns, the authors addressed instability issues in prior approaches ([62]) and demonstrated the framework's applicability across different sports contexts. Finally, with hierarchical clustering, similar movements were identified and grouped. While SMP provides an interesting solution to identify movements, it is not designed to create datasets to describe movements. Our framework builds on the discretization principles established by [64] but diverges by prioritizing dataset structuring: rather than extracting patterns from sequences, we segment continuous movement into discrete actions defined by velocity zones and aggregate these into events, creating stable, reusable datasets that support multiple analytical objectives.

Corbett et al. [20] introduced a change-point analysis methodology to iden-

tify natural transitions in physical output using binary segmentation applied to speed time series data from Australian Football. Rather than imposing fixed time windows, they automatically identified segments with distinct mean and variance properties based on statistical properties of speed signals. While this data-driven approach effectively captures speed transitions, it differs fundamentally from threshold-based segmentation: change-point analysis discovers segment boundaries through signal characteristics, whereas our framework uses predefined velocity zone thresholds to create semantically meaningful actions (standing, walking, jogging, etc.). This threshold-based approach enables direct comparison with prior Gaelic football research and facilitates interpretation by coaches and analysts who are familiar with velocity zone frameworks.

Adeyemo et al. [65] developed the LCCspm algorithm to extract l-length closed contiguous patterns from GNSS data in rugby league. Movement units were formed by discretizing velocity (walk, jog, run, sprint), acceleration (deceleration, neutral, acceleration), and turning angle (straight, acute-change, large-change, backwards) into 48 combinations encoded as characters. However, encoding movements into 48 character combinations increases complexity and reduces interpretability. In contrast, our framework deliberately uses velocity as the primary segmentation criterion for defining actions, with acceleration and turning angle treated as descriptive features rather than as segmentation variables. This simplification makes actions more interpretable and directly comparable to established velocity zone conventions in sports science.

The Sequential Movement Pattern framework was applied by Collins et al. [66] to extract movement patterns across competition levels in rugby league, identifying 121 unique patterns and using linear discriminant analysis to distinguish domestic and international competition. While effective for characterizing competition-level differences through pattern prevalence, their analytical pipeline focuses on which movement sequences are frequent in different contexts. Our framework shifts the emphasis to feature engineering: velocity zone discretization is used as a segmen-

tation strategy to create structured datasets where statistical features and running metrics are computed for each action and event.

In a recent study, Adeyemo et al. [67] compared three pattern mining algorithms (LCCspm, LCS, AprioriClose) for player position classification in rugby league, using movement unit discretization that combined velocity, acceleration, and turning angle into 48 combinations. Although this demonstrates that closed contiguous patterns carry positional information, the high-dimensional encoding (three descriptors per movement unit) makes individual patterns difficult to interpret semantically. Our research deliberately uses a lower-dimensional approach: velocity zone alone determines segmentation into actions, while acceleration and directional information are captured as statistical features (average, standard deviation, directional components) computed within each action.

2.1.4 Summary

Various approaches have been proposed in literature for transforming raw sensor data into features for sports analysis, including pattern mining algorithms and statistical feature extraction methods (summarized in Table 2.1). However, existing approaches lack detailed frameworks for structuring data into interpretable datasets that support diverse analysis. Moreover, while velocity zones are established in Gaelic football research, they are primarily used for descriptive analysis rather than as a systematic approach for dataset construction.

This dissertation in Chapter 3 addresses these gaps by introducing a systematic framework that transforms continuous GNSS and IMU data into two structured datasets: Action datasets (discrete velocity-zone segments) and Event datasets (aggregated activity periods). This framework uses established velocity zone thresholds for interpretability, provides explicit algorithms for reproducibility, preserves temporal sequences while enabling statistical analysis, integrates GNSS and IMU data, and supports multiple applications, including workload prediction and spatial analysis.

Table 2.1: Literature review of research in feature extraction from sensor data in sport.

Study	**Sport**	**Approach**	**Contribution**
Torres et al. [21]	Team sports	Narrative review of tracking systems	Establishes gap between technology deployment and analytical practice
Rampinini et al. [53]	Team sports	Velocity zone thresholds	Velocity zones definition for team sport
Malone et al. [23]	Gaelic football	4Hz GNSS; Descriptive statistics	Positional differences in running performance profiles
Malone et al. [19]	Gaelic football	4Hz GNSS; Descriptive statistics	Running performance changes across match phases
Ryan et al. [24]	Gaelic football	4Hz GNSS; Descriptive statistics	Acceleration profiles and positional demands
Cullen et al. [56]	Gaelic football	GNSS; Descriptive statistics	Activity profile in minor (U-18) players
Imbach et al. [57]	Football	GNSS with tsfresh Python module; statistical and frequency-domain features	Acceleration profile prediction
Midoglu et al. [59]	Women's soccer	Multi-modal integration; subjective wellness + GPS tracking	Large-scale dataset combining wellness and movement metrics

Study	**Sport**	**Approach**	**Contribution**
Hussain et al. [60]	Gym exercises	IMU wearable sensors; statistical and domain-specific feature extraction	Activity recognition and exercise classification
Fister et al. [61]	General sports	Minimalistic feature extraction toolbox	Efficient sensor feature extraction
Sweeting et al. [62]	Team sports	Spatio-temporal tracking; k-means clustering + Levenshtein distance	Automatic extraction of recurring movement sequences
White et al. [64]	Team sports	Sequential Movement Pattern mining; multi-descriptor discretisation	Improved movement discretisation and pattern identification
Corbet et al. [20]	Australian Football	Change-point analysis; binary segmentation on speed time series	Data-driven segment identification of intensity
Adeyemo et al. [65]	Rugby League	LCCspm algorithm; 48-combination multi-descriptor encoding	Closed contiguous movement pattern identification
Collins et al. [66]	Rugby League	Sequential Movement Pattern framework; linear discriminant analysis	Identification of movement patterns

Study	Sport	Approach	Contribution
Adeyemo et al. [67]	Rugby League	Pattern mining algorithms (LCCspm, LCS, AprioriClose)	Movement-based player position classification
Chapter 3 [49], [50]	**Gaelic football**	**GNSS + IMU; velocity zone-based segmentation**	**Framework to realize Action and Event datasets**

2.2 Spatio-Temporal Analysis in Team Sports

Having presented methods for structuring raw sensor data, the next challenge is extracting insights from player movements across space and time. In this section, graph-based approaches that reveal tactical patterns and spatial dynamics through network representations of team coordination are examined.

Graph-based representations provide a powerful framework for analyzing complex interactions and patterns in sports. By representing players or areas of the field as nodes and actions as edges, these methods enable insightful analyses of team strategies, player dynamics, and tactical decision-making. This section reviews relevant approaches to graph-based representations of movement and spatial dynamics, positioning the spatio-temporal graph framework presented in Chapter 4 within the existing literature and identifying its distinct contributions.

2.2.1 Graph Representations of Player Interactions

Graph-based representations have emerged as powerful tools for analyzing team sports, where nodes typically represent players and edges capture their interactions through passes, spatial proximity, or tactical relationships. These networks enable quantitative assessment of team connectivity, identification of key players through centrality metrics, and detection of tactical patterns through community structures. This section reviews studies that employ player-based graphs to analyze team dy-

namics and tactical organization, highlighting how our framework diverges by representing spatial areas rather than individual players as the fundamental units of analysis.

Raabe et al. [68] proposed Tactical Graph Networks that model spatio-temporal football data by representing players as nodes with weighted edges describing distances between them on the field. Their approach includes temporal connections linking the same player across consecutive time steps and demonstrates that graph-based representations can achieve performance comparable to state-of-the-art deep learning models while maintaining significantly lower complexity. The study incorporated various player attributes into node features and analyzed 15-second segments for binary classification tasks related to possession changes using multilayer Graph Neural Networks. Our research shares with this study a goal in modeling player movements within the field, but differs fundamentally in data sources, graph topology, and analytical purpose. While researchers in [68] focus on predicting possession-related outcomes, the spatio-temporal graph presented in Chapter 4 examines the spatial organization of movement by representing discrete areas of the field as nodes and movement transitions at different velocity zones as edges. This topology permits the identification of critical transition zones and spatial communities with distinct workload profiles that evolve across time windows, revealing insights about how the playing area is used by players rather than predicting discrete match events.

Gama et al. [69] applied network analysis to investigate intra-team activity during attacking phases of professional football, using betweenness centrality as a key metric to identify principal players who act as critical connectors in the team's attacking organization. Their approach constructs directed, weighted passing networks where players serve as nodes and edges represent ball transfers between them, with edge weights reflecting the frequency of passes performed or received. By analyzing these networks during attacking sequences, the authors identified which players occupied the most influential positions in facilitating ball progression and maintain-

ing attacking flow. The study demonstrated how graph-theoretic measures could reveal tactical roles and dependencies within a team's attacking structure, showing that certain players consistently emerged as key intermediaries for attacking moves. While the graph framework in Chapter 4 adopts similar centrality metrics to assess node importance, the graph topology diverges: it represents areas of the field as nodes rather than players, with edges capturing movements at different velocity zones rather than passes between teammates. This representation changes the analytical focus from identifying important players to assessing the significance of specific spatial zones in player movement patterns. Furthermore, it incorporates temporal dynamics through time windows that reveal how the importance of these zones evolves throughout the match, whereas in [69], the authors focus on aggregated attacking phases without examining the temporal evolution of spatial importance.

Silva et al. [70] developed uPATO (Ultimate Performance Analysis Tool), a software platform specifically designed for network analysis in team sports that generates adjacency matrices illustrating passing interactions among players within a team. The tool enables analysts to construct and visualize passing networks throughout a match or across distinct phases, computing various graph-theoretic metrics, including centrality measures (degree, betweenness, and closeness), clustering coefficients, and network density. uPATO's primary innovation lies in its capability to assess team performance over distinct time intervals within a match, allowing coaches and analysts to observe how passing patterns and player prominence evolve across different game phases. The software provides visual representations of passing networks where node size and positioning can reflect player importance and tactical roles, while edge thickness indicates passing frequency between players. This temporal segmentation enables the identification of how team connectivity and individual player roles change throughout the match. While both uPATO and the spatio-temporal graph presented in Chapter 4 use centrality scores and temporal segmentation, the approaches differ in terms of graph representation. Rather than

analyzing interpersonal passing dynamics among players, the spatio-temporal graph examines spatial movement patterns through discretized areas of the playing field, enabling the identification of the most traversed field cells and groups of cells that cannot be directly inferred from passing networks alone.

Gonçalves et al. [71] explored the relationship between passing networks, player positioning variables, and match outcomes in elite youth football. The study creates directed, weighted passing networks where players serve as nodes and edges represent passes between them, with various network metrics computed to characterize team connectivity and individual player roles. They calculated degree centrality (number of connections), betweenness centrality (extent to which a player acts as a bridge in passing sequences), closeness centrality (how quickly a player can reach all other teammates through passes), and clustering coefficient (tendency of players to form passing triangles). The research integrated these network measures with positional variables derived from GNSS tracking data, including player dispersion, team centroid position, and spatial distribution across the pitch. Their findings revealed that successful teams exhibited lower betweenness centrality scores, indicating less passing dependency on individual key players and more distributed ball circulation, and higher closeness centrality scores, suggesting well-connected passing relations where all players were readily accessible within the passing network. However, the authors focused primarily on static or aggregated positions throughout match phases rather than examining dynamic movements through space. The graph-based framework presented in Chapter 4 addresses a different task by focusing on how players move through discretized spatial cells over time rather than analyzing inter-player distance dynamics and their regularity. While [71] examined how players maintain spatial relationships with specific teammates, our work analyze how players traverse through different areas of the field at varying intensities, revealing which spatial zones serve as critical transition areas and how spatial communities with distinct workload profiles emerge.

2.2.2 Spatial and Tactical Analysis Methods

Researchers have developed different methods to analyze how teams organize themselves in space and how they implement tactics. These methods include analyzing player movement trajectories to understand coordinated patterns, dividing the field into zones to examine how space is used, and identifying tactical roles based on player positions. While these approaches provide valuable insights into team tactics and spatial behavior, they typically focus on specific events, single moments in time, or player positioning rather than tracking continuous movement across the field. This section reviews the main spatial analysis methods used in team sports research and thus provides context for capturing the way in which players traverse spatial zones at varying intensities over time.

Andrienko et al. [72] proposed an approach to analyze football player trajectories for extracting general patterns of coordinated movement using novel spatial and temporal reference systems. The method summarizes player positions by calculating their typical location and shows how much their actual positions varied from this typical spot. Connecting these typical positions creates a simplified movement pattern that shows how the team generally behaves. When tested on two Bundesliga games, the approach successfully detected meaningful patterns of team behavior defined by football experts. However, the evaluation remained limited to case studies, and practical application requires software tools that coaches and analysts can operate, as the current research prototype requires technical expertise to execute the complex query and aggregation procedures. While in [72] the authors' focus is on generalized collective trajectories at the team level, the spatio-temporal graph presented in Chapter 4 operates at a finer spatial granularity by dividing the playing area into discrete cells and constructing graphs that capture individual movement transitions. This granular representation permits identification of not only collective patterns but also localized critical zones.

Brefeld et al. [73] proposed a data-driven method to estimate individual movement models from positional data extracted from video recordings at 25 Hz. Their

approach analyzes player trajectories to construct probabilistic models of reachable positions given current position and velocity, creating zones of the field that each player can reach before opponents. The method divides player speeds into five ranges (standing < 0.28 m/s, walking 0.28–1.94 m/s, jogging 1.94–3.89 m/s, running 3.89–5.56 m/s, and sprinting > 5.56 m/s) and builds individual movement profiles using two alternative approaches: kernel density estimation for high accuracy and count-based histograms for real-time computational efficiency. These profiles represent the probability distribution of possible player positions within specific time horizons, demonstrating superior performance in predicting pass interception (97.5% accuracy) and pass completion (88.5% accuracy). This work relates to our research as both analyze spatial-temporal movement patterns from tracking data, though Brefeld et al. focus on modeling individual player capabilities to predict spatial control for tactical decisions, while our work examines collective movement patterns across discretized field areas to identify spatial communities characterized by distinct workload profiles.

In their study, Bialkowski et al. [74] developed a method for large-scale analysis of soccer matches using an entire season of spatio-temporal tracking data extracted from video recordings by Prozone (approximately 400 million data points from 375 games, sampled at 10Hz). They introduced a role-based representation that dynamically assigns each player to a tactical role at every frame, with roles discovered automatically using minimum entropy data partitioning (an approach that groups player positions to minimize spatial overlap between roles using probability density functions and clustering algorithms), effectively identifying formations and enabling both individual player and team analysis. Their method visualizes formation changes throughout matches using sliding time windows and tracks how individual players alternate between different tactical roles during games. This research aims to assign players to tactical roles based on their relative positions within team formations. On the contrary, our framework examines how players move through discretized spatial areas of the pitch and identifies spatial communities characterized

by distinct movement and workload patterns.

Fonseca et al. [75] applied Voronoi diagrams to reveal interaction patterns between attackers and defenders in Futsal, with players represented as nodes and boundaries between Voronoi regions serving as edges. The study found that attackers possessed larger and more regular dominant regions compared to defenders, with greater size variability among teammates. However, the analysis was limited to 19 plays within a specific sub-phase of Futsal in a reduced area. While Voronoi diagrams provide valuable insights into spatial dominance at specific moments, our framework extends beyond instantaneous configurations by tracking movements through discretized spatial cells over rolling time windows. This temporal dimension enables identification of not just which areas players dominate at a given moment but which spatial zones serve as critical transition points throughout the match, how spatial communities emerge and change, and how the physical demands associated with different areas evolve over time.

Machado et al. [76] present a visual analytics system for understanding match evolution through three techniques: Player Attribute Heatmaps showing where individual players are located, Tactical Scheme Heatmaps displaying defensive, midfield, or attacking formations, and pathline glyphs showing player trajectories from specific regions. Their approach treats the match as a sequence of time intervals and uses visualization to reveal when teams attack or defend. As opposed to their visualization approach, our work models the playing area as a spatial network and uses graph metrics (betweenness centrality and community detection) to identify how field zones functionally connect through movement intensity, revealing structural patterns of spatial organization rather than visualization tactical snapshots.

2.2.3 Spatio-Temporal Graph Models Beyond Sports

The following works, drawn from domains outside sports analytics, inspired the spatio-temporal graph framework developed in this thesis, demonstrating the versatility of graph-based representations for analysing movement and activity patterns

across diverse spatio-temporal contexts.

Vega-Oliveros et al. [77] presented two graph models to analyze fire activity in the Amazon basin (2003-2018): a static graph representing the entire period and a spatio-temporal graph representing weekly activity. The area of interest was divided into a grid with each cell representing a node. Two nodes were connected when successive events in chronological order occurred in their corresponding cells. This representation permitted analysis of the most affected areas (centrality score) and areas of similar activity (community detection) over different time windows.

Ferreira et al. [78] proposed Chronnet, a chronological network-based model for analyzing spatio-temporal data where nodes represent equal-sized areas obtained by dividing a geometric space into grid cells. Areas are connected when consecutive chronological events occur in distinct cells. Testing on three artificial datasets and a real dataset describing global fire activity (2003-2018) demonstrated the model's capability to extract data characteristics beyond basic statistics, including frequent patterns, spatial variations, and spatio-temporal clusters.

Jia et al. [79] introduced a frequency-based spatio-temporal graph network to analyze bike-sharing mobility patterns before (2019), during (2020), and after (2021-2022) COVID-19. Nodes represent locations in the area of interest, with weighted edges capturing the frequency of trips between locations. The analysis employed node connectivity, clustering coefficient, and modularity-based community detection to examine mobility pattern trends.

These works demonstrate the versatility of spatio-temporal graph representations across domains and establish a methodological foundation for our approach. Our framework adapts these principles to sports analytics by incorporating average speed as edge attributes, applying betweenness centrality to identify critical transition zones, and combining Louvain community detection with workload profiling to reveal spatial communities with distinct physical demand characteristics. This adaptation addresses the specific needs of sports performance analysis, where understanding not just where movement occurs but at what intensity and how spatial patterns evolve

is essential for tactical and conditioning insights.

2.2.4 Summary

The previous sections highlighted a growing use of graph-based methods for analyzing spatial and temporal dynamics in sports, but several gaps remain. Most passing network studies focus on ball events and player-to-player interactions rather than continuous movement through space. Spatial analysis approaches often provide static or snapshot-based views rather than capturing dynamic evolution. Furthermore, few approaches integrate spatial positioning, movement intensity, and temporal evolution within a unified framework. Table 2.2 provides a summary of this literature review.

The spatio-temporal graph framework presented in Chapter 4 addresses these gaps by representing spatial cells as nodes, encoding movements at different velocity zones as edge attributes, and analyzing these graphs across consecutive time windows. This representation enables identification of critical transition zones through betweenness centrality, detection of spatial communities representing areas of coordinated activity using Louvain community detection, characterization of these communities with distinct workload profiles based on speed and acceleration metrics, and examination of how these features evolve during the match. This integrated approach provides a comprehensive perspective on player behavior that extends beyond existing methods in sports analytics.

Table 2.2: Summary of literature review about spatio-temporal analysis in team sports and beyond.

Study	**Sport**	**Approach**	**Contribution**
Raabe et al. [68]	Football	Players as nodes; temporal GNN	Possession prediction (low complexity)

Study	**Sport**	**Approach**	**Contribution**
Gama et al. [69]	Football	Passing networks; betweenness centrality	Identifies key attacking players
Silva et al. [70]	Team sports	passing networks	Temporal evolution of player roles during game
Gonçalves et al. [71]	Youth football	Passing networks + spatial variables	Team connectivity linked to match success
Brefeld et al. [73]	Football	Individual movement control zones	97.5% pass interception; 88.5% completion accuracy
Andrienko et al. [72]	Football	Trajectory analysis; typical positions	Team behavior patterns
Bialkowski et al. [74]	Soccer	Probability density functions + clustering	Formation identification; tactical role evolution
Fonseca et al. [75]	Futsal	Voronoi diagrams	Attacker vs. defender spatial dominance
Vega-Oliveros et al. [77]	Fire activity	Grid cells as nodes; time windows	Hotspot identification; community detection
Machado et al. [76]	Soccer	Player Attribute Heatmaps; Tactical Scheme Heatmaps; pathline glyphs	Visual interpretation of match evolution and player positions
Ferreira et al. [78]	Fire activity	Chronnet; grid cells as nodes	Spatio-temporal patterns; clusters
Jia et al. [79]	Bike-sharing	Locations as nodes; frequency edges	Temporal network evolution

Study	Sport	Approach	Contribution
Chapter 4 [51]	**Gaelic football**	**Spatial cells as nodes; aggregated speed on edges; time windows**	**Critical zones, spatial communities, workload profiles, temporal evolution**

2.3 Metabolic Power and Workload Prediction

Spatial analysis reveals where and when players move, but it does not quantify energy demands. This section examines metabolic power frameworks for quantifying energetic demands and machine learning approaches for predicting athlete workload from sensor data. This aspect of the literature review examines the intersection of machine learning, wearable sensor technology, and athlete performance prediction in team sports, with particular focus on metabolic power assessment and workload prediction.

2.3.1 Metabolic Power: Theoretical Foundation

The metabolic power approach, originally developed by Di Prampero et al. [80], provides a framework for quantifying the energy demands of intermittent exercise by accounting for the metabolic cost of acceleration and deceleration. This foundational method addresses the limitations of velocity-based metrics alone, which fail to capture the energetic cost of frequent accelerations and decelerations characteristic of team sports. In their study, the authors established the biomechanical equivalence between accelerated running on flat terrain and constant-speed running on an incline, enabling calculation of metabolic power from GNSS-derived velocity and acceleration data. The model translates forward acceleration into an equivalent slope, from which energy cost can be derived using established relationships for uphill running. This theoretical framework assumes that the metabolic demands of accelerating horizontally are equivalent to running uphill at a constant speed, with

the slope determined by the magnitude of acceleration.

Osgnach et al. [81] introduced the practical application of metabolic power to elite soccer using GNSS technology. analyzing GNSS data from 56 Serie A matches, they demonstrated that metabolic power-based metrics (high metabolic load distance, equivalent distance) provided more comprehensive workload assessment than traditional speed-based metrics alone. The study found that football players covered on average $10,950 \pm 1,044$ meters per match, with an average energy expenditure of $61.12 \pm 6.57\,\mathrm{kJ \cdot kg^{-1}}$: 26% of the total distance was covered at high metabolic power (exceeding 20 W/kg), which accounted for approximately 42% of the total energy expenditure. This highlights that a substantial portion of metabolic load occurs during accelerations and decelerations, even at lower speeds, which traditional speed-based metrics fail to capture adequately. The study established metabolic power zones analogous to velocity zones, with high metabolic power defined as >20 W/kg.

The metabolic power framework has been updated in later publications [82], and [83]. In the first update, Di Prampero et al. [82] refined the estimation methodology to account for air resistance (contributing ≈2% of total energy expenditure) and to separate walking from running examples. At each time instant, the algorithm calculates the equivalent slope (ES) from the instantaneous velocity (v) and forward acceleration (a_f): $ES = a_f \cdot g^{-1} + k \cdot v^2 \cdot g^{-1}$, where g is gravitational acceleration, and $k \cdot v^2 \cdot g^{-1}$ the introduced air resistance component (not present in the original formula in [80]), with $k \approx 0.0037$ $\mathrm{N{\cdot}s^2{\cdot}kg^{-1}{\cdot}m^{-2}}$. Then, the transition speed v_{trans} (m/s) is then computed as a function of ES:

$$v_{trans} = -107.05 \cdot ES^5 + 113.13 \cdot ES^4 - 1.13 \cdot ES^3 - 15.84 \cdot ES^2 - 1.70 \cdot ES + 2.27. \tag{2.1}$$

If $v < v_{trans}$, the walking energy cost equation is used:

$$EC_w = (280.5 \cdot ES^5 - 58.7 \cdot ES^4 - 76.8 \cdot ES^3 + 51.9 \cdot ES^2 + 19.6 \cdot ES + 2.5) \cdot EM. \tag{2.2}$$

If $v \geq v_{trans}$, the running equation is used:

$$EC_r = (155.4 \cdot ES^5 - 30.4 \cdot ES^4 - 43.3 \cdot ES^3 + 46.3 \cdot ES^2 + 19.5 \cdot ES + 3.6) \cdot EM. \tag{2.3}$$

In EC_w and EC_r, EM represents the equivalent mass, which accounts for the additional load during acceleration: $EM = \sqrt{(1 + (\frac{a_f}{g})^2}$.

This refinement resulted in approximately 14% lower whole-match energy expenditure estimates compared to the original model in [80].

In later work, Osgnach et al.[83] demonstrated that the number of high-intensity efforts decreased throughout matches (from first to last 15-minute periods) without substantial changes in mean metabolic power (≈22 W/kg) or duration (≈6.5 s). However, mean metabolic power during low-intensity efforts decreased (5.8 to 4.8 W/kg), primarily due to longer recovery duration.

This dissertatino uses the original framework introduced in [80] to compute metabolic power from GNSS data, without the refinements proposed in [82]. This decision was made for the following reasons. First, to the best of the authors' knowledge, this work is the first attempt to predict metabolic workload transitions in team sports. For this reason, establishing a baseline using the widely adopted original formulation enables future comparisons and ensures reproducibility. In addition, recent studies evaluated the ecological validity of the metabolic power framework in Gaelic football [47], provided [48] and in hurling (another Gaelic field-invasive team sport) [84] used the original framework [80].

2.3.2 Metabolic Power in Gaelic football

Research into metabolic power assessment has changed the measurement of player workload in team sports, providing more accurate quantification of energetic demands than traditional velocity-based metrics. In Gaelic football, the adoption of metabolic power approaches has revealed profiles of physical activity and allowed for refined examination of positional and temporal differences in match demands. While these advances have considerably improved descriptive analytics, the literature is still dominated by retrospective analyses and lacks approaches for real-time, individualized workload prediction.

Malone et al. [48] examined the metabolic power and energetic demands of elite Gaelic football match play. Using 10Hz GNSS data from 50 elite players across 4 inter-county teams during 35 competitive matches (351 complete match samples), they provided the first comprehensive metabolic power profile of elite Gaelic football. Match mean distance was 9,222 $\pm$ 1,588 m with average metabolic power ranging from 9.5-12.5 W/kg and total energy expenditure of 58-70 $kJ \cdot kg^{-1}$. Significant positional differences emerged: midfielders showed the highest activity profiles (greater total distance, high-speed distance, high metabolic power distance, and energy expenditure) compared to other players. Temporal analysis revealed significant reductions in both metabolic power and traditional running variables as matches progressed. The rate of decline was comparable between metabolic power variables and speed-based variables. This study established that metabolic power is a valid and informative metric for physical demand, providing estimates of energetic expenditure within elite Gaelic football.

The ecological validity of metabolic power within elite Gaelic football was validated in a later publication by the same author. Malone et al.[47] examined the associations between metabolic power variables and objective aerobic fitness measures. Twenty-five elite players completed the following physiological testing at the start and at the middle of the season: VO_2max, velocity at VO_2max (vVO_2max), peak treadmill velocity, running economy, and blood lactate thresholds at 2 $mmol \cdot L^{-1}$

and 4 mmol·L^{-1}. Training and match-play running performance was continuously recorded using GNSS technology. Results demonstrated strong correlations between average metabolic power (r = 0.67-0.71), high metabolic load distance (r = 0.65-0.72), and aerobic fitness markers (VO_2max, vVO_2max). The study concluded that metabolic power variables offer reliable, real-time indicators of aerobic fitness responses to training and match demands in elite Gaelic football, suggesting their utility for monitoring adaptations throughout the season.

2.3.3 Machine Learning for Athlete Performance Prediction

Machine learning has become a valuable tool for predicting athlete performance in sports. By analyzing large amounts of sensor data, machine learning models can identify important patterns and forecast outcomes like physical decline or workload spikes. This section reviews how recent studies have used different data sources and techniques to predict athlete states and highlights the advances and limitations of these approaches for supporting decision-making in team sports.

A machine learning approach to analyze 50 professional soccer players from Serie A over a complete season has been proposed by Rossi et al.[85]. The study collected daily wellness questionnaires (7-point Likert scale for fatigue, sleep quality, stress, and muscle soreness), external workload features derived by GNSS data (total distance, high-speed running, accelerations, decelerations, and metabolic power variables) and RPE-based internal load. Using XGBoost with 10-fold cross-validation, they predicted next-day wellness status with high accuracy. Feature importance analysis revealed that sleep quality, muscle soreness, previous day's wellness, and specific GNSS variables (player load, total distance) were the most influential predictors. The study implemented both cross-validation and sequential scenarios (training on past data, testing on future). Although they include metabolic power as part of the input features, their main goal is to forecast wellness status for the next day, helping with weekly training planning rather than in-game player management. Our research differs because we predict objective workload transitions within matches,

using a metabolic workload ratio threshold personalized for each player. Instead of predicting how a player will feel the next morning, we predict metabolic work variations in short time windows, which helps in tactical decisions and substitutions during games.

Biro et al. [86] explored the applicability of machine learning methods to detect fatigue using sensor data from 19 athletes over repeated training sessions (steady runs and efforts). IMUs recorded triaxial acceleration, angular velocity, and magnetic orientation throughout repeated sessions. Raw time-series data were transformed into frequency and time-domain features, then used to train supervised ML models (Random Forest, Gradient Boosting Machines, and LSTM networks). The implemented model demonstrated high predictive accuracy for fatigue, showing significant correlations between predicted fatigue levels and observed performance declines. When comparing classifiers, they found that while LSTM networks showed advantages for sequential data, traditional machine learning algorithms (Random Forest) also achieved competitive performance, especially when augmented with well-engineered features. While in [86], the authors focus on controlled scenarios, predicting fatigue accumulation across sessions, in our work, presented in Chapter 5, the focus is on competitive match scenarios, predicting metabolic power transitions within ongoing games. Their IMU features are primarily frequency/time-domain transformations of accelerometer and gyroscope signals, while our features are composed of features used in sport literature describing distances, velocity zones, and cumulative workload. They predict absolute fatigue levels, while our study predicts binary transitions relative to player-specific baselines.

Cordeiro et al. [42] developed a machine learning-based framework to monitor and predict performance attenuation in Gaelic football players by integrating physiological, anthropometric, and subjective markers. The study included data from 41 male players, capturing neuromuscular metrics, perceptual responses (fatigue, mood, sleep quality), biochemical markers (creatine kinase levels), and anthropometric variables (body mass, fat percentage). Performance attenuation was calcu-

lated by measuring the change in each athlete's physical and biochemical metrics from before to after match play. The difference for each metric (post-match minus pre-match) quantified the degree of performance decline due to match demands, ranking athletes from minimal to strong attenuation. These rankings served as the predictive labels for the machine learning models, allowing them to forecast which players were likely to experience higher fatigue and reduced physical performance after matches. The dataset was balanced using multiple synthetic oversampling techniques (SMOTE, ROSE, ADASYN) to address class imbalances between different levels of performance attenuation. Various machine learning algorithms, including Random Forest, AdaBoost, and XGBoost, were evaluated, with Random Forest combined with ADASYN achieved the highest F1-score (0.75), affirming the importance of strength, lower-body power, and speed/agility as key predictors. There are, however, important differences between this research and ours. In [42], authors focus on predicting declines in performance using data from before and after the match. In contrast, our research predicts shorter-term workload transitions, specifically whether a player will enter an elevated metabolic workload in the next 3 minutes, based on rolling in-game features. Furthermore, while in [42] authors analyze data for each match as a whole, our research uses short time windows to track each player's workload during the game.

In their study, Dijkhuis et al. [87] analyzed tracking data from 302 Dutch Eredivisie soccer matches during the 2018–2019 season to develop machine learning models that predict players' physical performance early in the match, providing decision support for substitutions. Players' physical performance was defined using three progressively sensitive variables: total distance covered, distance covered within speed categories, and energy expenditure within metabolic power categories. The authors defined five speed categories and five power categories, normalizing player performance against individual baselines. These variables were calculated for fixed time intervals, including five-minute windows, 15-minute periods, halves, and entire matches. The physical performance of individual players in these intervals was

normalized relative to their personal baseline values, computed as the mean performance over the entire season and specifically for each 5-minute period of a match. Players were classified as underperforming if they failed to reach 100%, 95%, or 90% of their individualized baseline in the given interval. Features used consisted of absolute and relative metrics, such as percentage of distance covered, percentage of distance covered in each speed category relative to corresponding baseline averages, and percentage of energy expenditure in each power category compared to baseline. These were calculated in both current and cumulative forms up to the specific time window. Machine learning models, including Random Forest, Decision Tree, and a Naive Bayes baseline, were trained and evaluated using these features. Training employed SMOTE to balance the dataset for underperforming and performing classes. The Random Forest model using energy expenditure in power categories achieved the best predictive performance across all underperformance thresholds. In contrast to Dijkhuis et al., whose goal is to predict early (using the first 15 minutes of data) if a player will maintain performance levels across the entire remaining match, our approach provides continuous predictions throughout the game. This permits real-time monitoring and intervention opportunities, which differ from a single early assessment.

In addition, their analysis does not explore the temporal dynamics of performance within sessions: by aggregating session data for prediction, their work may overlook crucial intra-session variations (for example, periods of acute high demand or rapid recovery), which are important for monitoring player performance. By focusing specifically on dynamic, time-window based workload transitions rather than aggregated session averages, our study advances the field's capacity to detect changes in physical demands within matches. In addition, authors in [88] do not explicitly handle the interpretability of their machine learning models, which can impact trust in decision-making from coaches or sport scientists. Our work includes interpretable methods with feature importance analyzes to provide clear explanations of how input variables influence model predictions, enabling coaches and practitioners to

understand which factors most contribute to workload transitions.

Morciano et al. [88] aimed to classify above-average performance in professional football players using supervised machine learning algorithms trained on biometric parameters obtained from wearable sensors. Data were collected from 63 training sessions of 20 professional players, stratified by role (strikers, midfielders, and defenders). Wearable sensors included GNSS and IMU at 50 Hz, capturing accelerometers, gyroscopes, magnetometers, and heart rate data. Four biometric parameters served as features for the classification: energetic cost, metabolic power, anaerobic threshold, and maximum oxygen uptake rate (VO2max). Player performance was assessed through seven performance indicators (acceleration distance, deceleration distance, high-intensity distance, and distances in speed ranges 16-20 km/h, 20-24 km/h, and >24 km/h), each calculated as session-aggregated averages. The classification task aimed to predict if each performance indicator was above or below the team average for that training session. Random Forest models, optimized through grid search and whale-optimization algorithms, achieved accuracy exceeding 90% across all player roles. In their work, Morciano et al. used biometric parameters collected during training sessions to predict whether players achieve above-average session performance. On the contrary, our approach uses real-time movement data to predict imminent metabolic work rate transitions during matches. Morciano's method classifies overall session performance, whereas our method forecasts dynamic, upcoming changes for real-time intervention

2.3.4 Summary

The reviewed literature (summarized in Table 2.3) establishes metabolic power as a valid framework for quantifying energetic demands in team sports and demonstrates the potential of machine learning for athlete monitoring, though predictive approaches have primarily targeted wellness states, aggregate performance metrics, or post-match performance decrements. However, existing studies do not provide frameworks for predicting imminent workload transitions within matches using short

temporal windows, limiting the ability to anticipate workload changes that could inform tactical decisions such as substitution timing. Our study, presented in Chapter 5, addresses these gaps by introducing a novel metric based on metabolic power capturing work-recovery balance; implementing player-specific thresholds; predicting short-term (3-minute) transitions rather than match-level outcomes; combining current-window and cumulative features to capture both acute activity and chronic fatigue effects; and demonstrating that accurate metabolic workload transition predictions are achievable. This provides the foundation for future operational systems when appropriate technical infrastructure is in place.

Table 2.3: Comparison of metabolic power assessment and workload prediction approaches in team sports.

Study	**Sport**	**Input Data**	**Time Scale**	**Experiment Design**	**Prediction Target**	**Approach**
Di Prampero et al. [80]	Running	Lab experiments	Theoretical	Biomechanical modeling	Energy cost framework	Theoretical model
Osgnach et al. [89]	Soccer	GNSS 10 Hz	Match-level	Descriptive analysis	Metabolic power profile	Descriptive statistics
Malone et al. [48]	Gaelic football	GNSS 10 Hz	Match-level	Descriptive analysis	Metabolic power by position	Descriptive statistics
Malone et al. [47]	Gaelic football	GNSS 10 Hz + physio	Match-level	Validation study	Metabolic power vs. VO2max	Correlation analysis
Rossi et al. [85]	Soccer	GNSS + wellness	Daily	Previous day → next day	Wellness classification	Classification
Biro et al. [86]	Multi-sport	IMU	Session-level	Within-session aggregated	Fatigue classification	Classification
Dijkhuis et al. [87]	Soccer	Tracking	Early prediction	First 15 min → full match	Under performance detection	Classification
Cordeiro et al. [42]	Gaelic football	GNSS, IMU, physio	Pre/post-match	Match response analysis	Performance attenuation ranking	Classification
Morciano et al. [88]	Football	GNSS 50 Hz, IMU 50 Hz, HR	Session-aggregated	End-of-session classification	Above/below-average performance	Classification
Chapter 5	**Gaelic football**	**GNSS 10 Hz + IMU 100 Hz**	**Rolling 3-min windows**	**Continuous intra-match**	**Metabolic Workload transitions**	**Classification**

2.4 Automatic Detection of Game Phases in Sport

Workload prediction would benefit from game context: for example, whether the team is attacking or defending. Section 2.4 reviews automatic event detection methods that classify game phases from sensor data, providing necessary context without

manual video annotation.

Manually labelling game events remains a major challenge in professional sports analysis. Sports analysts must rewatch matches and use specialized software to mark events like possession changes, attacking phases, and shot attempts with precise timestamps. This manual labelling provides valuable context for understanding the physical performance patterns from sensor data. However, this process is time-consuming, expensive, and prone to errors.

This section reviews methods for automatically detecting events and game phases in sports, using tracking data, video analysis, and wearable sensors to identify important moments without manual work

Richly et al. [90] develop an automated approach for identifying compound events in football by analyzing positional data. They try to demonstrate the feasibility of detecting key events such as passes, receptions, shots on target, and clearances directly from spatio-temporal player and ball tracking data using supervised machine learning. The data used in this work consisted of professional football matches from the German Bundesliga, tracked using optical sensors that provided two-dimensional coordinates for players and the ball. The authors manually created a gold standard from video recordings of game sections. The methodology developed involved computing distinct feature types from raw positional data, such as velocity, acceleration, acceleration peaks (both maximum and minimum), direction change (computed as the angle between consecutive direction vectors), distance to target (Euclidean distance from an object's current position to the assigned goal), and a cross on target line feature (distance from the goal to the point where an object's trajectory would cross the goal line). These features were computed for the ball, and their values at each timestamp formed the basis for event classification. Three supervised machine learning algorithms were trained and compared: Support Vector Machines (SVM), K-Nearest Neighbors (KNN) with k=3, and Random Forest. SVM and Random Forest achieved recall between 59.9-76.7% and precision between 35.1-38.7%, except for reception events where precision dropped to 5.9-8.9%. Important

limitations of their work included the small training dataset relative to the diversity of football match scenarios. Furthermore, the ball-centric feature set may not capture all relevant information from the movements of the players. Our research, presented in Chapter 6, employs a team-centric approach using pairwise inter-player movement relationships and focuses specifically on detecting game phases ending in shots. Finally, our primary innovation lies in demonstrating that a single team's wearable sensor data is sufficient to infer shot events.

Link et al. [91] propose computational models for detecting individual and team ball possession in soccer based on positional tracking data from 60 German Bundesliga matches in the 2013/14 season. Their approach classifies ball possession into various hierarchical types, such as Individual Ball Possession (IBP), Individual Ball Action (IBA), Individual Ball Control (IBC), Team Ball Possession (TBP), Team Ball Control (TBC), and Team Playmaking (TPM). The data consisted of player and ball xy-coordinates collected at 25 Hz using an optical tracking system. The authors developed a three-step detection process: (1) preprocessing to remove tracking errors, (2) identifying IBP/IBA boundaries using distance, velocity, and acceleration between player and ball, and (3) classifying ball control using Bayesian networks. Our approach differs from their framework by demonstrating that a specific game phase can be inferred from team movement coordination patterns alone, without the ball and opponents' position.

Khaustov et al. [92] present a rule-based algorithm designed to reconstruct key soccer events (ball possession, successful and unsuccessful passes, and shots on goal) from spatiotemporal player and ball positional tracking data. Their work uses two distinct datasets comprising continuous or episodic sequences of positional coordinates for all field players and the ball. They extract four main feature types derived from spatiotemporal trajectories: player speed and direction, ball velocity metrics, and ball acceleration characteristics. Characteristic temporal and spatial patterns matching each event type are encoded as rules to identify and classify events effectively. Validation on available datasets confirms high accuracy and computational

efficiency. However, the methodology in [92] requires full tracking data (both teams and ball coordinates) and pre-encoded knowledge about event signatures that may not transfer across sports. In contrast, our approach uses supervised machine learning on engineered pairwise player movement features derived from wearable data, requiring no data about the other team and ball's position. This allows the model to learn sport-specific and context-specific patterns directly from data.

Vidal-Codina et al. [93] developed a deterministic two-step algorithm to automatically extract football events using player and ball tracking data. The approach presents a decision tree-based algorithm that determines which player owned the ball at each frame by analyzing spatial proximity and ball velocity and identifying distinct player configurations during intervals when the ball is not in play (corners, free kicks, throw-ins). The event detection step then combines changes in ball possession to classify in-game events (passes, shots, crosses, saves, etc.). When evaluated across multiple football tournaments and tracking data providers, the algorithm achieved detection rates of 90% or higher for most event categories, with particularly high accuracy for shot classification. In contrast to Vidal-Codina et al., our approach works without explicit ball position coordinates at each second and data from both teams. While their approach achieves superior accuracy (90%+), it requires expensive video tracking infrastructure unavailable to most teams globally.

The study by Bischofberger et al. [94] presents a validation of a simple rule-based algorithm designed to detect passes and shots from positional data across four diverse datasets. The methodology involves detecting playing events based on thresholds of ball acceleration and player proximity, followed by classifying these events using rule-based rules and machine learning models such as decision trees and ensemble classifiers. They extract features such as distances to the goal, opening angle to the goal, initial kick speed, whether the receiver is a goalkeeper or outfield player, and directional metrics. These interpretable features serve as inputs to rule-based classifiers and machine learning models to distinguish shots from passes. The study only focused on two event types (passes and shots), excluding more complex

or subjective events like tackles, dribbles, or fouls. Pass detection achieves high F-scores up to 0.92 when positional and event data are dependent, but accuracy drops significantly (to around 0.71 for passes and 0.65 for shots) when positional and event data are independent, signaling a real-world challenge in event temporal extraction. This is due to the fact that for certain datasets, positional data might have been post-processed or reconstructed using event annotations. For shot classification, they achieved F-scores between 0.83 to 0.91 with their rule-based classifiers and up to 0.95 with machine learning classifiers.

Unlike Bischofberger et al., who work with synchronized high-quality tracking data and manual annotations, our work uses incomplete wearable sensor data where tracking is available only for one team. Moreover, contextual data are available only for specific types of playing events, and the exact position of the ball is unknown. Despite this more challenging situation, we achieve competitive shots classification performance through novel pairwise feature engineering that captures inter-player coordination patterns rather than explicit positional information.

2.4.1 Summary

The previous section described various approaches to automatic event and phase detection in team sports, summarized in Table 2.4. These approaches vary from rule-based algorithms to deep learning models. Existing approaches rely on ball position, both teams' sensor data, or video footage. These methods achieve high accuracy but require expensive infrastructure and complete tracking systems that are unavailable to most teams. As a result, existing studies have not addressed the practical scenario faced by many clubs: detecting game phases using only wearable sensor data from a single team without ball position or opponent sensor data. Indeed, only clubs with enough resources can use advanced contextual analysis tools.

Our study, presented in Chapter 6, addresses this gap by demonstrating that key game phases (specifically shooting events) can be automatically classified using only wearable sensor data from one team through a novel pairwise feature engineering

approach, without requiring ball position or opponent tracking data. This method achieves competitive classification performance (F1 scores 0.67-0.69). Furthermore, the framework is transferable and can be applied to other invasion-field team sports beyond Gaelic football.

Table 2.4: Comparison of automatic event and game phase detection approaches in team sports.

Study	**Sport**	**Data Type**	**Data Coverage**	**Events Detected**	**Detection Method**	**Feature Source**
Richly et al. [90]	Football	Optical tracking (stationary cameras)	Both teams + ball (x, y)	Passes, receptions, shots, clearances	Supervised ML (SVM, KNN, Random Forest)	Ball-centric (velocity, acceleration, direction, distance to goal)
Link et al. [91]	Soccer	Optical tracking system	Both teams + ball	Individual and team ball possession, ball control, playmaking	Bayesian networks	Distance, velocity, acceleration between player-ball
Khaustov et al. [92]	Soccer	Tracking datasets	Both teams + ball	Ball possession, passes, shots	Rule-based algorithm	Player speed, direction, ball velocity, ball acceleration
Vidal-Codina et al. [93]	Football	Video footage	Both teams + ball	Passes, shots, crosses, saves	Decision tree algorithm	Spatial proximity, ball velocity, player configurations
Bischofberger et al. [94]	Football	Tracking datasets	Both teams + ball	Passes, shots	Rule-based + ML (decision trees, ensemble)	Ball acceleration, player proximity, goal distance, opening angle
Chapter 6 [52]	**Gaelic football**	**Wearable sensors**	**Single team only**	**Shooting phases (both teams)**	**Supervised ML + Recursive Feature Elimination**	**Pairwise inter-player movement differences (adjacency matrices)**

2.5 Literature Review Summary

This chapter identified critical gaps across four research domains that this thesis addresses through novel methodological contributions.

1. **Feature Engineering Gap** (Section 2.1): While wearable sensors have become ubiquitous in team sports, the literature lacks systematic frameworks for transforming raw sensor data into structured datasets. Existing work focuses on what features to extract (velocity zones, accelerations) but not how to algorithmically organize data to enable diverse analyses. Chapter 3 addresses this

gap by introducing a multi-step framework to realize the Action and Event datasets, which permit analytical tasks, while reducing dimension with respect to raw sensor data but maintaining essential information integrity.

2. **Spatial Analysis Gap** (Section 2.2): Existing graph approaches in team sports model player-to-player interactions through passing or event networks, treating players as nodes. This structure captures tactical relationships but cannot reveal how teams collectively move through space. This remains critical for understanding territorial control and transition dynamics. Chapter 4 approaches spatial analysis by representing spatial cells (not players) as graph nodes, with edges encoding movement transitions between locations. This cell-based approach enables the identification of critical spatial zones, community detection of regions with similar movement characteristics, and temporal analysis of how spatial patterns evolve. These insights are inaccessible to player-centric models.

3. **Workload Prediction Gap** (Section 2.3): Metabolic power provides more comprehensive energetic demand quantification than velocity-based metrics, with validated ecological validity in Gaelic football. While recent work has begun exploring in-game performance prediction, these approaches share a limitation: they predict aggregate outcomes (will the player underperform the full match? Is session performance above average?) rather than imminent transitions (will workload change in the next 3 minutes?). Chapter 5 addresses this gap by introducing a player-specific Metabolic Work Ratio metric quantifying work-recovery balance, and rolling 3-minute window predictions of imminent workload transitions rather than match-aggregate outcomes. This enables proactive in-match intervention when real-time data streams and proper technology are operationally available.

4. **Identification of Game Phases Gap** (Section 2.4): the literature review demonstrates that all prior automatic playing event detection methods require

ball position data, both teams' tracking, or video footage. This requires expensive resources that may be unavailable to most clubs. No prior work addresses event detection using only single-team wearable sensors. Chapter 6 fills this critical practical gap by classifying shooting events for both teams using pairwise inter-player movement features derived exclusively from one team's sensor data.

The following chapters present detailed methodologies and experimental validations designed to address the gaps identified in this chapter.

Chapter 3

Developing a Threshold-Based Activity Feature Set

The work presented in this chapter has been published in:

1. Antonini Valerio, Alessandra Mileo, and Mark Roantree. "Engineering features from raw sensor data to analyse player movements during competition." Sensors 24.4 (2024): 1308.
2. Antonini Valerio, Dermot Sheridan, and Mark Roantree. "A multidimensional database of in-game player movements (Actions and events) in gaelic football." Data in Brief 57 (2024): 111082.

3.1 Introduction

The previous chapter provided an overview of wearable sensor data and their characteristics in invasion field team sports such as Gaelic football. This chapter addresses an important practical gap in sports analytics. While wearable sensors combining GNSS and IMU technologies are widely used to capture high-frequency player movement data, many studies and commercial applications do not fully disclose transparent and reproducible data processing pipelines. Prior research in the analysis of players movement in Gaelic football using sensor data identified running

and acceleration profiles [23], [24], studied the relationship between technical performance indicators and running performance [95], or investigated activity profile, player load, and heart rate [54], typically focusing on GNSS data alone. However, there remains a gap for comprehensive and reproducible frameworks that effectively integrate both GNSS and high-frequency IMU data into structured datasets, capturing more detailed movement dynamics. To address this need, this chapter presents a methodological framework composed of algorithms that convert raw GNSS and IMU data into datasets reflecting player movement intensity described by summary metrics of the variables returned by sensors. These algorithms are fully documented with pseudocode to promote transparency and facilitate adaptation in both research and applied settings, with a particular focus on Gaelic football.

The main contribution of the framework is a segmentation algorithm that partitions continuous sensor data into continuous movement segmented by velocity thresholds and an aggregation algorithm that groups these units into higher-level sequences framed by periods of rest or low-intensity movement. They provide reproducible, systematic steps to transform raw sensor inputs into structured datasets.

The primary objective of this work is to demonstrate the framework can organize wearable sensor data to preserve established analytical capabilities while enabling additional forms of analysis that can be challenging to perform directly from raw sensor data. For example, by structuring data into movement units and contextual sequences, the framework facilitates examination of velocity zone transitions, duration patterns, and proportion of high-speed distance between resting or low-intensity states. To validate the framework, two complementary case studies are presented. The first shows that the derived datasets produce results consistent with established Gaelic football research. In particular, the Action dataset preserves the same analytical information of the raw sensor data while drastically reducing the dimension of the data. The second illustrates the practical application of the structured datasets in a predictive modelling context, forecasting high-speed running distance in the second half of matches. While this prediction can be made using raw sensor data, our

framework creates clear and organized datasets that make it easier to understand and analyze player movement.

In summary, this work contributes to the methodology of sports analytics by providing both an algorithmic framework and resulting structured datasets. This dual-dataset approach provides a foundation for future research, offering systematic data organization that supports a wide range of analytical objectives.

3.2 Raw Sensor Data

This section provides a comprehensive overview of the wearable sensor data collected during 11 competitive inter-county Gaelic football games. First, the data collection process is described, followed by a presentation of the structure and characteristics of the raw data.

3.2.1 Data Collection

Data were collected during 11 competitive Gaelic football inter-county games during the years 2019, 2020, and 2021 on the same team. The data collection was part of different research: the football club shared access to STATSports cloud-based Sonra platform with the university for the analysis. During those games, players wore a micro GNSS sensor device (STATSports Apex 10 Hz, Northern Ireland, UK), placed on the upper back, within a vest that fitted the body comfortably. The accuracy and reliability of the specific device have been independently validated [96]. Beato et al. reported that the 10 Hz Apex unit produced a small distance bias of approximately 1–2% across 400-m, 128.5-m circuit, and 20-m linear running trials, and a peak speed bias of 2.36 $\pm$ 1.67% against a radar gun criterion, with a nearly perfect intraclass correlation coefficient of 0.96 [96]. These findings support the use of the device for measuring both distance and velocity variables during training and match play. It should be noted, however, that these validation trials were conducted under optimal, controlled conditions involving isolated, linear running bouts on

an outdoor athletic track, with a high satellite count and minimal environmental interference. As such, accuracy estimates derived from such protocols may not fully generalize to the continuous, multidirectional demands of competitive intercounty Gaelic football match play, where movement patterns, intensity fluctuations, and environmental conditions (stadium vs outdoor athletic track) differ substantially [46, 96]. Nonetheless, subsequent interunit reliability analysis of the same 10 Hz Apex unit during a standardized intermittent running protocol demonstrated excellent reliability for total distance (ICC = 0.999) and good reliability for peak speed (ICC = 0.898) [97], providing further confidence in the consistency of data collected across multiple units within the same session.

Data were collected during 11 competitive Gaelic football inter-county games during the years 2019, 2020, and 2021 on the same team. The number of games reflects the constraints of an opportunistic data-sharing arrangement with an elite county team, which is common in applied sports science research where access to elite populations is governed by institutional agreements. The games cover three seasons and different stages of the championship calendar, introducing natural variation in conditions across fixtures. All players involved were members of the registered senior inter-county squad selected by management for each fixture, ensuring the data reflect genuine elite competition. It should be noted that the 2020 season was played without spectators due to Covid-19 restrictions, though the physical demands captured by the sensors are unlikely to have been systematically affected by crowd absence.

The 37 players included in the dataset were not selected by the researcher but represent all squad members who participated in at least one of the 11 games and for whom valid sensor data were available. No exclusion criteria were applied at the data collection stage based on position, fitness level, or playing time, ensuring the dataset reflects the full composition of an elite inter-county squad.

3.2.2 Data Structure and Preprocessing

The GNSS sensor records data at 10 Hz, providing 10 data points per second for latitude, longitude, and speed (m/s). In addition, the built-in inertial measurement unit captures accelerometer and gyroscope data at 100 Hz. This means that for every second of play, 100 data points are captured from each IMU channel. The accelerometer axes represent the forward-backward (anteroposterior), side-to-side (mediolateral), and vertical (up-down) directions of movement, respectively, while the gyroscope captures rotational motion about these axes.

Data were downloaded using Sonra STATSports software (version 4.5.19). For each session (training or game), raw data were downloaded in CSV format. The software provides the option to download raw or summary data for the entire session or divided by drills (training) or by quarters or half of games. For each game, a separate CSV file was extracted containing the raw data for each player for each quarter of the game, with all files related to a single game stored within one folder. In this way, the data were returned already filtered and cleaned from the software: for example, periods such as the warm-up and the interval between the halves were not included. The raw data exported were pre-filtered by the proprietary company to reduce the noise within the GNSS signal [98]. In most cases, the software only provided data for those players who were actively involved in the playing area, while returning an empty CSV file for players on the bench in that quarter. All players who participated in any portion of a game were retained in the dataset at the data collection stage, including those who played only a limited number of minutes. Contextual filtering was then applied prior to each individual analysis depending on the specific requirements of that analysis (for example, excluding players who played five minutes or less, or restricting analyses to players who completed the entire game). The filtering operation is later specified before each analysis. For some games, the software returned data for players who were not involved in the game. Those players have been removed by comparing sensor data with official match reports. The additional data pre-processing performed was to exclude data

points where the recorded speed exceeded 10 m/s or instantaneous acceleration was greater than $\pm$ 6 m/s^2, as suggested by previous research [99]. These values were replaced with the value from the previous timestamp. Sonra software offers the option to extract metrics used in sport research (for example, total distance, sprints, max speed, and many others). The raw data were downloaded, and key metrics were computed using thresholds or formulas established in sport science literature. This approach also provides the flexibility to manipulate the data according to the specific task.

It should be acknowledged as a limitation that the data were collected across the 2019, 2020, and 2021 intercounty seasons, a period during which Gaelic football underwent notable rule changes that may have altered the physical demands of competition. Furthermore, the 2020 season was played behind closed doors due to Covid-19 restrictions, which may have influenced tactical communication and player intensity profiles. The 2019–2021 window was not chosen by experimental design but reflects the period during which the institutional data-sharing agreement between the county team and the university was active. As such, direct longitudinal comparisons across seasons should be interpreted with caution.

A sample of the downloaded raw data, extracted using the STATSports Sonra software, is presented in Table 3.1. The full dataset contains 76,313,030 rows, reflecting the high sampling rates of the sensors. Each row represents a unique observation taken every 1/100th of a second (0.01 second). Within each second, there are 100 rows capturing accelerometer and gyroscope measurements (Accl X, Y, Z and Gyro X, Y, Z), which vary from one row to another. Meanwhile, position (latitude, longitude) and speed observations updated every 10th row (at 10 Hz), meaning these values remain the same across 10 consecutive rows before changing. Player and Game IDs displayed have been randomly generated to maintain anonymity.

Figure 3.1 shows 30 seconds of player speed sampled at 10 Hz for Player 153 during Game 934. Fluctuations in the speed signal indicate the player alternated between phases of faster running and periods of reduced speed or recovery.

Table 3.1: Sample of raw wearable sensor data downloaded using Sonra software. Data were recorded at 10 Hz for time, latitude, and longitude, and at 100 Hz for accelerometer (Accl X, Y, Z) and gyroscope (Gyro X, Y, Z) measurements.

Player	Time	Half	GameID	Lat	Lon	Speed	Accl X	Accl Y	Accl Z	Gyro X	Gyro Y	Gyro Z
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.415	0.635	0.074	-47.46	113.05	-9.53
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.406	0.449	0.097	-69.79	110.81	-3.57
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.326	0.273	0.114	-96.11	105.28	2.45
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.368	0.102	0.119	-119.21	108.01	5.60
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.344	0.012	0.111	-137.48	107.73	7.63
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.300	0.010	0.105	-120.50	108.20	6.00
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.335	0.025	0.100	-110.20	111.00	5.20
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.320	0.030	0.090	-115.00	107.85	4.80
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.310	0.040	0.089	-130.00	104.00	4.55
220	14:18:19.1	2	893	53.21	-6.15	3.5	0.315	0.045	0.088	-125.00	105.50	4.90
220	14:18:19.2	2	893	53.22	-6.14	3.6	0.310	0.038	0.087	-120.00	106.00	5.00

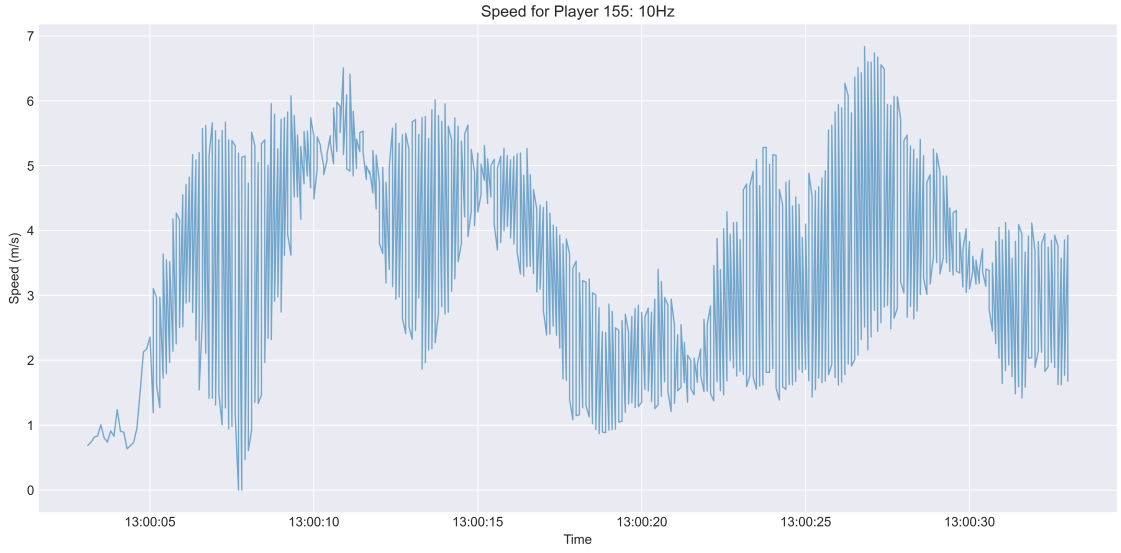


Figure 3.1: 30 seconds of speed data sampled at 10 Hz for Player 153 during Game 934

Figure 3.2 presents 30 seconds of raw tri-axial accelerometer data (Accl X, Accl Y, Accl Z) sampled at 100 Hz for Player 152 during Game 934. Each subplot displays the fluctuations in acceleration along a different axis. The Accl X and Accl Y axes show frequent spikes and oscillations, which may reflect forward and backward movements (Accl X), lateral movements (Accl Y), and changes of direction or impacts. The Accl Z axis generally oscillates around zero but displays both upward and downward spikes, indicating vertical movement.

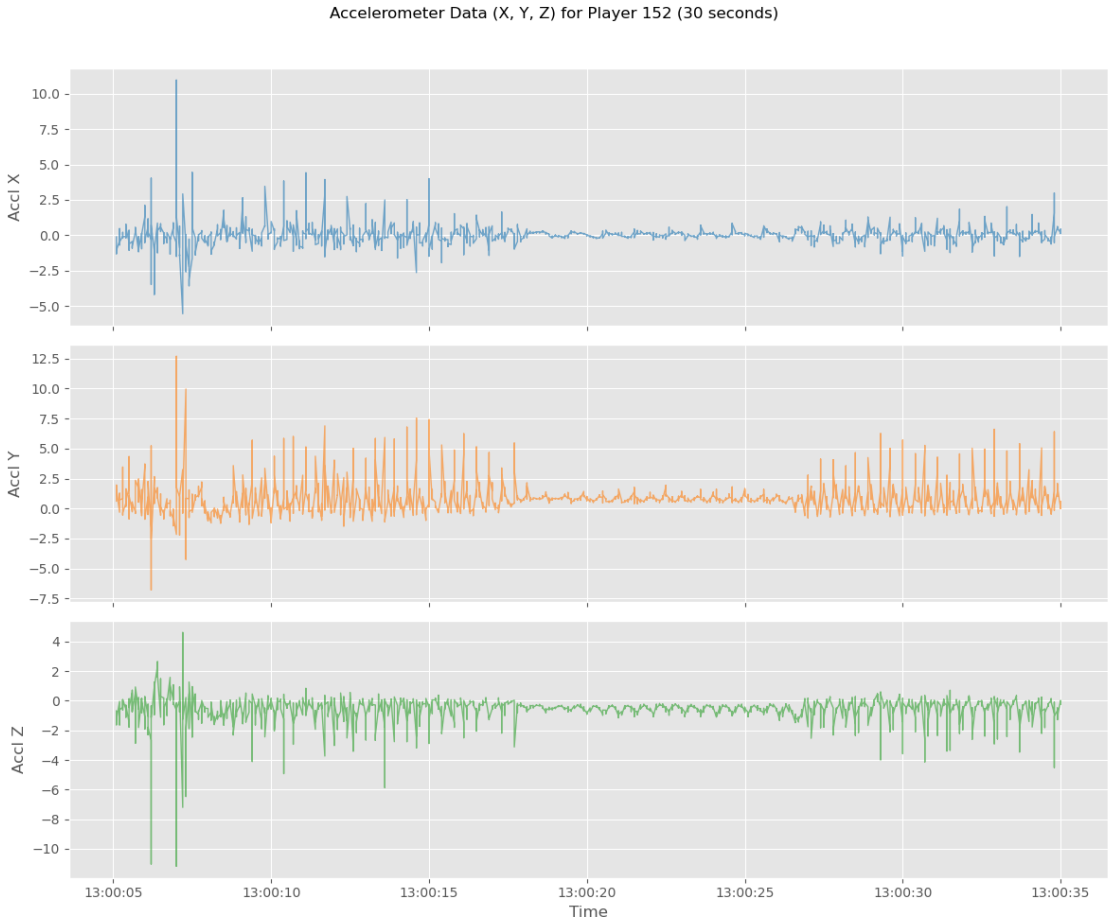


Figure 3.2: 30 seconds of tri-axial accelerometer data sampled at 100 Hz for Player 152 during Game 934

3.2.3 Overall Dataset Characteristics

Table 3.2 provides a comprehensive summary of the dataset characteristics. The dataset includes 11 competitive inter-county Gaelic football games played between 2019 and 2021, involving 37 unique players across 7,647 minutes of match play. The

high-frequency sampling (10 Hz for GNSS, 100 Hz for IMU) resulted in over 76 million data rows. Performance metrics show considerable variation across games, with total distances ranging from 8,352 to 9,208 meters (m) per player, high-speed distances at speed $\geq$ 4.72 m/s) from 1,402 to 1,744 m, and sprint distance ($\geq$ 6.11 m/s) between 441 and 556 m. As discussed in Section 2.1.1, different velocity thresholds have been used in Gaelic football literature. In this section, high-speed distance (distance at speed $\geq$ 4.72 m/s) and sprint distance (distance at speed $\geq$ 6.11 m/s) are computed using velocity thresholds defined by [19]. The use of these thresholds is to ease comparison with running profiles defined in literature ([23] and [19]). Later in this chapter and in subsequent chapters, different velocity thresholds are used, based on [53] and later applied in Gaelic football [54].

Table 3.2: Summary of dataset scope and characteristics

Characteristic	**Value**
Number of games	11
Years covered	2019, 2020, 2021
Number of players	37
Total minutes of play	7,647
Total data rows	76,313,030
GNSS sampling rate	10 Hz
IMU sampling rate	100 Hz
Average game duration	80.1 $\pm$ 1.8 min
Players per game (complete)	11.0 $\pm$ 1.6
Performance Ranges (per player)	
Total distance	8,352 - 9,208 m
Distance $\geq$ 4.72 m/s	1,402 - 1,744 m
Distance $\geq$ 6.11 m/s	441 - 556 m
Max speed	8.2 - 8.7 m/s

Table 3.3 shows the distribution of speed values in the entire dataset (every player included, regardless of the role or the minutes played): the average speed during match play is 1.97 m/s per player, with a standard deviation of 1.51 m/s. The distribution of player speed is right-skewed, as evidenced by the mean speed being greater than the median (1.50 m/s). This right skewness means that while most player movements happen at moderate speeds, there is a longer tail of higher speed values caused by intense activity such as sprints or rapid accelerations. This

reflects the intermittent nature of Gaelic football, where players mostly move at lower speeds, with occasional moments at higher speeds.

Table 3.3: Descriptive statistics for player speed (m/s) from the raw data.

Statistic	Speed (m/s)
Mean	1.97
Standard Deviation	1.51
25th Percentile	0.97
Median (50th Percentile)	1.50
75th Percentile	2.85
Maximum	9.92

The boxplots in Figure 3.3 reveal that while median player speeds remain close across halves, there is a slight decrease in median speed during the 2nd half compared to the 1st half. The maximum typical speed (excluding outliers) indicated in the plot by the upper whisker is slightly lower in the second half compared to the first half.

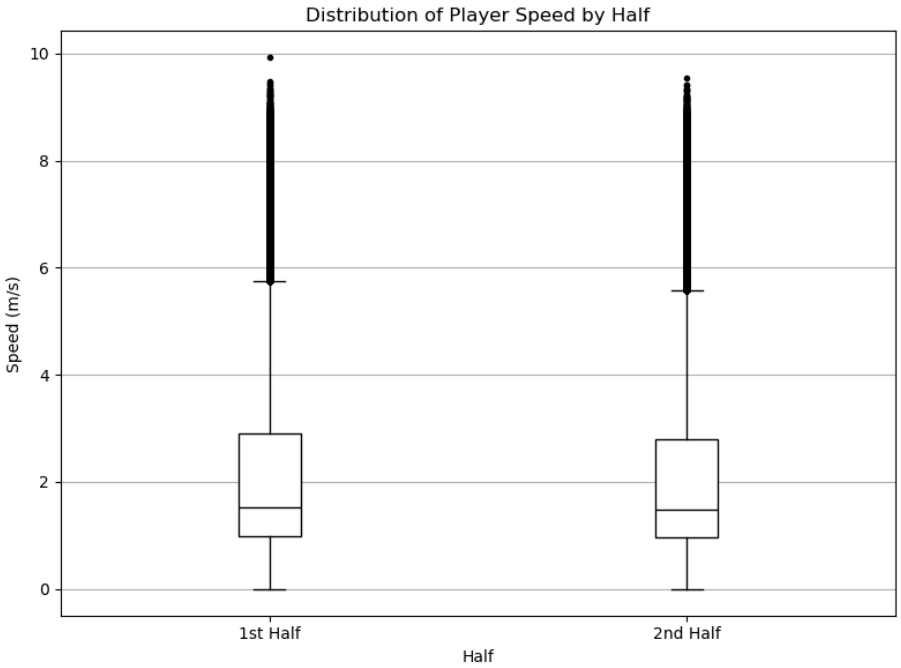


Figure 3.3: Box-plot distribution of speed by half of the game.

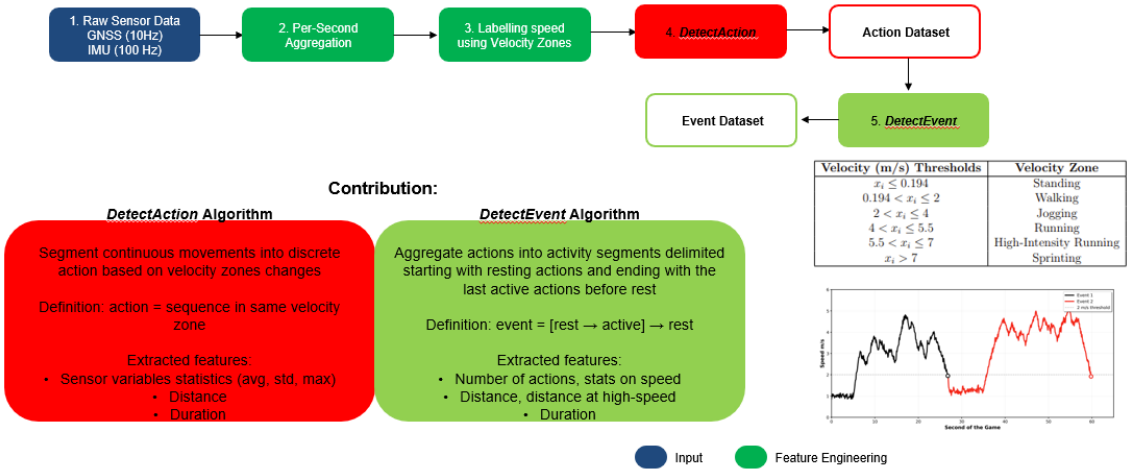


Figure 3.4: Overview of the multi-step framework for transforming raw GNSS and IMU sensor data into the Action and Event datasets. The pipeline progresses from raw sensor data through per-second aggregation, velocity zone labelling, and the DetectAction and DetectEvent algorithms, producing two structured datasets that preserve player movement information.

3.3 A Multi-Step Framework for Running Activity Datasets

In this section, the multistep feature engineering-based framework is presented. The framework, illustrated in Figure 3.4, takes as input the raw data described in the previous section and returns two datasets describing the running activity of players:

- An **Action** dataset. Each row describes an action performed by a player during one game. In this research, an *action* is defined as any uninterrupted sequence of seconds occurring in the same velocity zone. Velocity (or speed) zones in team sports are labels of running speeds identified by threshold values and used to categorize athlete movement during training or competition [100]. These velocity thresholds can be determined arbitrarily by the proprietary software of the tracking device or expressed relative to a physiological test [100]. Formally, an action is a temporal interval $[t_s, t_e]$ during which a player's velocity remains within the same velocity zone, defined as:

$$A = \{t \in [t_s, t_e] : Z(v(t)) = z\}$$

where t represents time in seconds, $v(t)$ is the player's speed at time t, $Z(\cdot)$ is the velocity zone labeling function that maps speed values to the velocity zones standing, walking, jogging, running, high-intensity running, and sprinting according to the velocity thresholds in Table 3.4, z is a constant velocity zone, and $t_s < t_e$. The velocity thresholds defining the velocity zones suggested by [53] (Table 3.4), and later applied in Gaelic football [54], were used.

The Action dataset is defined as

$$\mathcal{A} = \{A_i = (t^{(i)}_{\text{start}}, t^{(i)}_{\text{end}}, z_i)\}_{i=1}^{N},$$

where each A_i is an action characterized by the contiguous time interval $[t^{(i)}_{\text{start}}, t^{(i)}_{\text{end}}]$ and a velocity zone label $z_i \in V$.

Table 3.4: Velocity thresholds and corresponding velocity zones used to label player movement intensity, as defined by Rampinini et al. [53] and applied to Gaelic football in this thesis. Each threshold defines the speed range (x_i, in m/s) associated with a discrete movement category, from standing to sprinting.

Velocity (m/s) Thresholds	**Velocity Zone**
$x_i \leq 0.194$	Standing
$0.194 < x_i \leq 2$	Walking
$2 < x_i \leq 4$	Jogging
$4 < x_i \leq 5.5$	Running
$5.5 < x_i \leq 7$	High-Intensity Running
$x_i > 7$	Sprinting

- An **Event** dataset. Each row describes an event performed by a player during a game. In this research, an *event* is defined as a sequence of actions that begins when a player transitions from a resting state (standing or walking at speed ≤ 2 m/s) to an active state (jogging, running, high-intensity running, or sprinting) and ends when the player returns to a resting state.

  Formally, first we need to divide the velocity zones into two disjoint subsets corresponding to resting and active states:

$$V_{\text{rest}} = \{\text{standing}, \text{walking}\} \quad (\text{speeds} \leq 2 \text{ m/s}),$$

$$V_{\text{active}} = \{\text{jogging}, \text{running}, \text{high-intensity running}, \text{sprinting}\}.$$

For each game and player, the ordered sequence of actions represented by

$$\mathcal{A} = \{A_i = (t_{\text{start}}^{(i)}, t_{\text{end}}^{(i)}, z_i)\}_{i=1}^{N},$$

where each action A_i has a time interval $[t_{\text{start}}^{(i)}, t_{\text{end}}^{(i)}]$ and a velocity zone label $z_i \in V = V_{\text{rest}} \cup V_{\text{active}}$.

An event

$$E_j = (A_{i_j}, A_{i_j+1}, \ldots, A_{k_j})$$

is defined as a sequence of actions defined by the following properties:

1. **Event start:**

   $$z_{i_j} \in V_{\text{active}},$$

   and either

   $$i_j = 1, \quad \text{or} \quad z_{i_j-1} \in V_{\text{rest}},$$

   meaning that the event begins at the start of data or with the first active action following a resting action.

2. **Event continuation:** For all intermediate indices,

   $$z_m \in V_{\text{active}}, \quad \text{for } m = i_j, \ldots, k_j - 1,$$

   meaning the player stays in active velocity zones throughout the event.

3. **Event end:**

   $$k_j = N, \quad \text{or} \quad z_{k_j+1} \in V_{\text{rest}},$$

   or

   $$z_{k_j} \in V_{\text{active}},$$

the event ends just before the player returns to a resting state ($z_{k_j+1} \in V_{\text{rest}}$) or at the end of the dataset ($k_j = N$). In summary, each event captures a continuous period of activity bounded by rest periods.

The Event dataset is the collection

$$\mathcal{E} = \{E_1, E_2, \ldots, E_M\},$$

partitioning the entire sequence $\mathcal{A}$ into consecutive events. Each event E_j is characterized by its aggregated features, including player ID, event start time $t_{\text{start}}^{(i_j)}$, event end time $t_{\text{end}}^{(k_j)}$, the sequence of contained actions, and others described later in the chapter.

3.3.1 From Wearable Sensor Data to the Action dataset

The framework implemented to obtain the Action dataset from wearable devices data, describing the running activity level of players, is described in detail in the following steps.

1. **Data Acquisition.** The acquisition of raw data is explained in Section 3.2.1. The wearable sensor worn by players is composed of a 10Hz GNSS and a 100Hz IMU device. The GNSS device records 10 observations for each second for the variables latitude, longitude, and speed (m/s), while the IMU records 100 observations for each second for the variable acceleration over the x, y, and z axes. A sample of the input raw data of the framework is shown in Table 3.1.

2. **Step 1. Per-Second Aggregation.** For each player, the wearable sensor used to acquire the data returns 10 values each second of the variables latitude, longitude, and speed. While some analyses may benefit from such high-level granularity, this work aggregates the data into one value per second. The downsample of 10Hz sensor data to a rate of 1Hz, by computing the average of the variables every 10 samples, is a method already used in sport

Table 3.5: Sample of the aggregated data. The names of some columns in the header have been shortened: latitude (Lat.) and longitude (Lon.).

Player ID	**Time**	**Lat.**	**Lon.**	**Speed (m/s)**	**Acceleration (m/s^2)**
152	15:49:51	54.11398	−7.21719	5.36	0.22
152	15:49:52	54.99321	−7.49778	5.60	0.24
152	15:49:53	54.11334	−7.89912	5.14	-0.46
152	15:49:54	54.34509	−7.21196	5.11	-0.03

science research [101]. High-frequency sensor data capture detailed variations within one second but also include noise and volume that may not be relevant to the analysis. Downsampling to 1 Hz reduces data volume and noise, improving computational efficiency without sacrificing important information (for the scope of this research). It should be acknowledged, however, that averaging within-second speed values may attenuate the detection of very short high-intensity bursts characteristic of Gaelic football, where sub-second accelerations and brief sprinting efforts may be partially smoothed by aggregation. This represents a known trade-off between temporal resolution and computational tractability, and findings derived from the 1 Hz representation should be interpreted with this in mind.

The input dataset of this step is the raw data, where each row is 1/100 of a second, where acceleration x, y, and z have different values for each row, while the 10Hz GNSS variables have a different value every 10 rows (Table 3.1). The output is represented by the temporal roll-up of 1 value per second (Table 3.5). The location (latitude and longitude) is updated to the centroid of the locations covered during the second (average of 10 subsequent locations), while the speed would be the average value during the second. The acceleration is computed by subtracting each speed value from the previous one and adding it as a new column. In the same way as for the 10Hz features, the 100Hz acceleration x, y, and z are downsampled to 1Hz by computing the average of the 100 data points for each second.

3. **Step 2. Per-Second Labeling.** Each speed value x_i ($i = 1, 2, .., n$, where n is the last second of the game) at time i is labeled into velocity zones using

Table 3.6: Sample of the aggregated data with labeled actions. The names of some columns in the header have been shortened: latitude (Lat.), longitude (Lon.), velocity zone (V. Zone), and acceleration (Acc. (m^2/s)).

Player ID	Time	Lat.	Lon.	Speed (m/s)	V. Zone	Acc. (m^2/s)
152	15:52:11	54.12311	−7.35743	5.42	Running	0.04
152	15:52:12	54.11387	−7.19723	3.97	Jogging	−1.45
152	15:52:13	54.23334	−7.29989	5.10	Running	1.12
152	15:52:14	54.21387	−7.13780	5.07	Running	−0.03
152	15:52:15	54.11390	−7.48109	4.99	Running	−0.08

thresholds indicated in Table 3.4. Therefore, for each second of the game, speed values are converted to one of the six velocity zones: 'standing', 'walking', 'jogging', 'running', 'high-intensity running', and 'sprinting'.

A sample of the dataset after this step is shown in Table 3.6.

4. **Step 3. Detect and Aggregate Action.** In this work, an action is defined as any uninterrupted sequence of seconds occurring in the same velocity zone. The minimum duration of an action is 1 second if the player enters that specific velocity zone only for 1 second, changing zone in the successive second. The 1s interval dataset exposed in Table 3.6 now serves as input to this step, where the goal is to further aggregate the data: in this case, a roll-up to the level of action. Each data point will no longer represent 1 s of activity but an entire action, which will likely run over a number of seconds. This is carried out using the proposed DetectAction algorithm described in Algorithm 3.1. DetectAction algorithm takes as input the 1-second dataset in Table 3.6 divided by players and half of the game (as it requires data to be consecutive), processes the rows of the dataset, compresses similar labels into a unified row, and returns an Action dataset where each row is a specific action described by a set of features. The entire list of features describing the action is displayed in Figure 3.7.

   Algorithm 3.1 iterates through the N rows of a dataset (Table 3.6) comprising the sequence of the time, location, speed, action, and acceleration of a player during a game. At each iteration i, the action performed by a player is indicated by the variable x_i and the time when the action occurs by the variable z_i.

Table 3.7: Features extracted from each action in the Action dataset. For each discrete action summary statistics are computed from GNSS-derived speed and IMU-derived tri-axial acceleration signals, alongside spatial and temporal descriptors.

Feature Name	**Description**
Avg Speed	Average speed of the player during the action interval (m/s)
Min Speed	Minimum speed reached during the action (m/s)
Max Speed	Maximum speed reached during the action (m/s)
Std Speed	Standard deviation of the speed, measuring variability (m/s)
Avg Acceleration	Mean magnitude of acceleration over the action (m/s^2)
Max Acceleration	Maximum acceleration magnitude during the action (m/s^2)
Min Acceleration	Minimum acceleration magnitude during the action (m/s^2)
Std Acceleration	Standard deviation of acceleration magnitude
Avg Accel X	Average acceleration along the X-axis (m/s^2)
Max Accel X	Maximum acceleration along the X-axis (m/s^2)
Min Accel X	Minimum acceleration along the X-axis (m/s^2)
Std Accel X	Standard deviation of acceleration in X-axis
Avg Accel Y	Average acceleration along the Y-axis (m/s^2)
Max Accel Y	Maximum acceleration along the Y-axis (m/s^2)
Min Accel Y	Minimum acceleration along the Y-axis (m/s^2)
Std Accel Y	Standard deviation of acceleration in Y-axis
Avg Accel Z	Average acceleration along the Z-axis (m/s^2)
Max Accel Z	Maximum acceleration along the Z-axis (m/s^2)
Min Accel Z	Minimum acceleration along the Z-axis (m/s^2)
Std Accel Z	Standard deviation of acceleration in Z-axis
Duration	Total duration of the action interval (seconds)
Distance	Total distance covered during the action (metres)
High Acc Events	Count of high acceleration occurrences above a defined threshold during the action
Total Distance	Cumulative distance covered during the action

Table 3.8: Sample of the aggregated data after the compression of similar actions. The names of some columns in the header have been shortened: start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), average speed (Avg. Speed (m/s)), and velocity zone (V.Zone).

Player ID	Start Time	End Time	Start Lat.	Start Lon.	End Lat.	End Lon.	Avg. Speed (m/s)	V. Zone	Duration
152	15:53:51	15:53:55	54.123	−7.357	54.224	−7.351	5.36	Running	4
152	15:53:56	15:53:58	54.224	−7.351	54.011	−7.391	3.97	Jogging	2
152	15:53:59	15:54:04	54.011	−7.391	54.349	−7.650	4.98	Running	5
152	15:54:05	15:54:11	54.349	−7.650	54.012	−7.655	3.83	Jogging	6
152	15:54:12	15:54:14	54.012	−7.655	54.020	−7.657	4.51	Running	4

The first step of the algorithm is to initialize the required variables avg_speed, avg_speed_list, start_time_list, end_time_list, y, and z, which will measure the average speed during the action, the list of such speeds, the time when the action begins and the time when it ends, the current stored action, and the row when the action starts, respectively (1–6). While iterating over each row of the dataset (which has a length equal to N) (7), if the current performed action is equal to the current stored action, then it continues with the next row (8, 9). If the two actions are different (10), then the average speed from the row when the action starts to the current row is computed (11) and stored in the list of the speeds (12), the starting time and ending time of the action are added to the corresponding lists (13, 14), and the variables storing the row when the action starts (z) and the current action (y) are updated (15, 16). If the row is the last row of the dataset (17), the algorithm computes the average speed until this row (18), it stores the value in the list of the speeds (19), and it adds the starting time and ending time of the action to the corresponding lists (20, 21).

Algorithm 3.1 exclusively displays the variables average speed, starting time, and ending time. The variables recording the starting and ending locations adhere to the rule of the starting and ending times, while the remaining extracted features (Table 3.7) follow the criterion of average speed. Table 3.8 displays a sample of the resulting Action dataset returned by DetectAction algorithm, with some of the extracted features omitted for space reasons.

A key distinction between the Action dataset and conventional GNSS-derived

Algorithm 3.1 DetectAction Algorithm

Require: N rows of dataset
Require: $N \geq 2$
Require: x_z action performed by the player at iteration z
Require: t_z time of the game at iteration z

$avg\_speed \leftarrow 0$
$avg\_speed\_list \leftarrow []$
$start\_time\_list \leftarrow []$
$end\_time\_list \leftarrow []$
$y \leftarrow current\_action$
$z \leftarrow 0$
while $i \leq N$ **do**
    if $x_i = y$ **then**
        continue
    else if $x_i \neq y$ **then**
        $avg\_speed \leftarrow avg(x_z, .., x_{i-1})$
        $avg\_speed\_list[z] \leftarrow avg\_speed$
        $start\_time\_list[z] \leftarrow t_z$
        $end\_time\_list[z] \leftarrow t_{i-1}$
        $z \leftarrow z + 1$
        $y \leftarrow x\_i$
    else if $i = N$ **then**
        $avg\_speed \leftarrow avg(x\_z, .., x\_i)$
        $avg\_speed\_list[z] \leftarrow avg\_speed$
        $start\_time\_list[z] \leftarrow t_z$
        $end\_time\_list[z] \leftarrow t_i$
    end if
end while

summary metrics needs explicit statement here. Commercial GPS software typically returns pre-computed metrics (for example total distance, high-speed running distance, sprint counts) aggregated over fixed temporal windows such as quarters, halves, or full games. In the present framework, the unit of analysis is not a fixed time window but the individual action, defined as a contiguous sequence of seconds during which a player operates within the same velocity zone. Every feature in Table 3.7 is computed at the level of the action rather than the session period, preserving the temporal granularity and sequential structure of player movement in a way that aggregate metrics cannot.

This design, however, introduces a methodological dependency that must be acknowledged as a limitation. Because action boundaries are determined entirely by the velocity zone classification applied in Step 2, the validity of every action-level feature subsequently extracted is directly contingent on the accuracy of the GNSS-derived velocity signal. Errors in velocity measurement do not merely distort a summary metric computed on top of a fixed window: they propagate into the segmentation itself, potentially misclassifying action boundaries, corrupting action duration estimates, and altering the sequential structure of the data on which downstream analyses depend. This represents a more fundamental form of measurement-induced uncertainty than is present in conventional fixed-window approaches. As discussed in Section 3.2.1, the STATSports Apex 10 Hz unit demonstrated a peak speed bias of $2.36 \pm 1.67\%$ under controlled validation conditions [96], and accuracy is known to decrease during high-intensity, multidirectional activity [46]. These known limitations of the GNSS velocity signal should be considered when interpreting action-level findings throughout this thesis.

5. **Step 4. Direction of Action.** Based on a formula already used in past research in sport science [64], it is possible to exploit consecutive pairs of latitude and longitude to identify the direction of movement of each action. The direction of an action is given by the turning angle, which represents the

angle that a player turns to reach the ending from the starting location. To obtain the turning angle, we first need to compute the bearing, which measures the angle between two consecutive points, measured clockwise from magnetic north [64]. In other words, it represents the direction a player has to travel in a straight line from the starting point to the ending point, measured in degrees. The following represent the steps needed to compute the bearing and then the turning angle:

(a) Compute X and Y [64]:

$$X = \cos(\theta B)\sin(\Delta L), \tag{3.1}$$

$$Y = \cos(\theta A)\sin(\theta B) - \sin(\theta A)\cos(\theta B)\cos(\Delta L),, \tag{3.2}$$

where θ denotes the latitude, L denotes the longitude, A and B denotes two consecutive decimal coordinates, and ΔL denotes the difference between L at A and L at B.

(b) Compute the bearing β [64]:

$$\beta = \arctan 2(X, Y), \tag{3.3}$$

(c) The turning angle τ is obtained by subtracting the bearing computed on two successive time points [64]:

$$\tau = \beta_i - \beta_{i-1}, \tag{3.4}$$

The turning angle values are expressed as degrees from 0° to 360°, and then are converted to a label (forward, backward, right, and left), indicating the direction of movement. The thresholds and labels for the turning angle values are shown in Table 3.9.

Table 3.9: Direction of movement expressed as labeled turning angles.

Turning Angle (Degrees)	**Direction**	**Direction ID**
$45 \leq \tau_i$ or $\tau_i \geq 315$	Forward	1
$\tau_i > 45$ and $\tau_i < 135$	Left	2
$\tau_i \geq 225$ and $\tau_i < 315$	Right	3
$\tau_i \geq 135$ and $\tau_i < 225$	Backward	4

6. **Step 5. Action Dataset Description.**

   Table 3.10 displays the Action dataset, which represents one of the two output datasets returned by the proposed multi-step framework. Each row of the dataset represents an action performed by a player during gameplay. Each action can have a different duration, with a minimum duration of one second. The extracted features describe the action, enriching the info in terms of speed, acceleration, and direction.

Table 3.10: Sample of the action data after the compression of similar actions. The names of some columns in the header have been shortened: start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), average speed (Avg. Speed (m/s)), velocity zone (V. Zone), and ID of the direction of movement (Dir. ID). Some of the extracted features are omitted due to space reasons.

Player ID	Start Time	End Time	Start Lat.	Start Lon.	End Lat.	End Lon.	Avg. Speed (m/s)	V. Zone	Duration	Dir. ID
152	15:53:51	15:53:55	54.123	−7.357	54.224	−7.351	5.36	Running	4	1
152	15:53:56	15:53:58	54.224	−7.351	54.011	−7.391	3.97	Jogging	2	1
152	15:53:59	15:54:04	54.011	−7.391	54.349	−7.650	4.98	Running	5	2
152	15:54:05	15:54:11	54.349	−7.650	54.012	−7.655	3.83	Jogging	6	1
152	15:54:12	15:54:14	54.012	−7.655	54.020	−7.657	4.51	Running	4	3

3.3.2 Aggregating Action Data into Event-Level Data

This section describes the steps required to aggregate the actions into events and construct the Event dataset.

1. **Step 6. Event Detection.** In this research, an event is defined as a sequence of one or more actions that begins immediately after a resting state and ends with the last active action before the player returns to a resting state. Resting states are defined as standing or walking actions (speed ≤ 2 m/s). In detail, an event starts when the player transitions from a standing or walking action to perform any action at higher speed (jogging, running, high-intensity running,

or sprinting) and ends with the last such active action before the next standing or walking action is encountered.

It is important to clarify that transitions between active velocity zones do not terminate an event and do not require passage through a resting state. For example, a player who transitions directly from a jogging action into a high-intensity running or sprinting action, without returning to standing or walking, remains within the same event throughout. Each such velocity zone occupied during the event constitutes a distinct action within it, captured separately in the Action dataset. The event therefore represents the full uninterrupted bout of active movement, which may internally contain actions spanning multiple active velocity zones, from jogging through to sprinting, before the player returns to a resting state.

For each detected event, a set of summary features is extracted and stored in the Event dataset. These include event duration (s), total distance covered (m), high-speed running distance (m at speed ≥ 5.5 m/s), the number and types of constituent actions, and further descriptors of the speed and acceleration profile across the event. This feature set enables quantitative characterisation and comparative analysis of events across players, and games.

As illustrated in Figure 3.5, which shows player speed over time as a visual representation:

- **Event 1** (black line): Begins at the start of data, includes initial resting actions (speed ≤ 2 m/s), transitions to active actions (speed > 2 m/s), and ends when a resting action (standing or walking) is encountered.
- **Event 2** (red line): Begins immediately after the standing or walking action that ended Event 1, continues through additional resting actions, then transitions to active actions, and ends at the next standing or walking action.

The figure uses speed values to visualize the activity patterns, where the 2 m/s

threshold corresponds to the boundary between resting and active actions as defined in Table 3.4.

It should be acknowledged as a limitation that the event boundary definition, by design, uses only standing and walking as termination conditions. This means that an event may internally encompass a wide range of movement intensities (from low-intensity jogging through to sprinting) without a new event boundary being created. This design follows the work-rest interval paradigm established in team sport science [46], where rest is operationally defined as low-speed locomotion and is consistent with how effort bouts have been defined in Gaelic football research [54]. However, future work could explore alternative segmentation strategies.

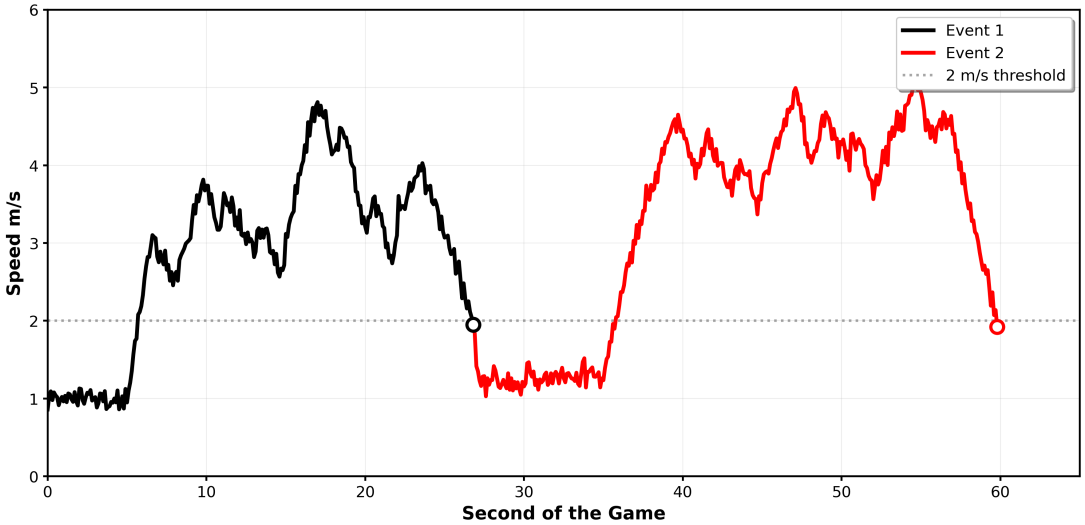


Figure 3.5: Sequential event detection visualized through speed data.

The DetectEvent algorithm proposed in Algorithm 3.2 iterates through the N rows of the dataset corresponding to one half of a player's game (Table 3.10) to identify sequences of actions defining events.

Algorithm 3.2 begins with the initialization of variables: y_s, y_w, events_list, id_event, and current_state, representing a variable equal to 'standing', a variable equal to 'walking', the list of the events ids, the current event id, and the current state, respectively (1–5). Current state is an indicator that contains 0 the first time the algorithm meets an action equal to 'standing' or 'walking'

Algorithm 3.2 DetectEvent Algorithm

Require: $N \geq 2$
Require: x_z action performed by the player at the iteration z
1: $y_s \leftarrow$ *'Standing'*
2: $y_w \leftarrow$ *'Walking'*
3: $events\_list \leftarrow []$
4: $id\_event \leftarrow 0$
5: $current\_state \leftarrow 0$
6: **while** $i \leq N$ **do**
7: **if** $x_i = y_s$ or $x_i = y_w$ **then**
8: **if** $current\_state = 1$ **then**
9: $id\_event \leftarrow id\_event + 1$
10: $events\_list[i] \leftarrow id\_event$
11: **end if** $current\_state = 0$
12: $events\_list[i] \leftarrow id\_event$
13: $current\_state \leftarrow 1$
14: **end if**$x_i \neq y_s$ and $x_i \neq y_w$
15: $events\_list[i] \leftarrow id\_event$
16: $current\_state \leftarrow 0$
17: **end while**

or 1 if it has already met one. While iterating over the dataset (6), if the current action x_i is equal to 'standing' or 'walking' (7), two situations can occur. If the current state is equal to 1 (8), a new event starts: the id_event is increased by 1 (9), and the new value is added to the list of events (10). On the contrary, if the current state is equal to 0 (11), add the id_event to the list of the events (12) and switch the value of the variable current state to 1 (13). If the current action x_i is different from 'standing' or 'walking' (14), add the current id_event to the list of the events (15) and set the current state to 0 (16).

At the end of this step, each action is assigned to an event with a unique identifying label (Table 3.11).

2. **Step 7. Event Dataset Description**. Once the events are identified by the DetectEvent algorithm, the following step is the aggregation of actions of the same event. The Action dataset is aggregated into the Event dataset by grouping rows sharing the same event ID. As already done with the actions, for each event, features describing the level of activity (Figure 3.12) were

Table 3.11: Sample of the Action dataset after event detection. The names of some columns in the header have been shortened: player ID (P.ID), start second (Start Sec.), end second (End Sec.), start latitude (Start Lat.), start longitude (Start Lon.), end latitude (End Lat.), end longitude (End Lon.), velocity zone (V. Zone), direction (Dir.), duration (Dur.), distance (Dist.), and event ID (E.ID). Some of the columns have been omitted due to space reasons.

P.ID	Start Sec.	End Sec.	Start Lat.	Start Lon.	End Lat.	End Lon.	Avg. Speed	V. Zone	Dir.	Dur.	Dist.	E.ID
152	54	56	54.55	−7.35	54.11	−7.35	5.36	Running	1	4	20.3	31
152	56	57	54.11	−7.35	54.13	−7.36	6.60	H.I. Running	1	1	6.6	31
152	57	62	54.13	−7.36	54.33	−7.37	4.98	Running	3	4	17	31
152	62	71	54.33	−7.37	54.52	−7.34	3.83	Jogging	2	9	18.8	32
152	71	76	54.51	−7.34	54.55	−7.33	4.51	Running	4	6	24.5	32

extracted. Each row of the final Event dataset represents an event performed by a player in a game.

Table 3.13 displays a sample of the resulting Events dataset. 'Player ID' indicates the ID of the player; the 'Event ID' indicates the ID of the event; 'Start_Second' and 'End_Second' indicate the starting and ending seconds (expressed as second of game); 'Start_Lat', 'Start_Lon', 'End_Lat', and 'End_Lon' indicate the starting and ending locations of the event; 'Avg Speed' and 'Std Speed' indicate the average and the standard deviation of the speed of the actions composing the event; 'Duration' indicates the duration of the event; 'Std Duration' indicates the standard deviation of the duration of the actions composing the event; 'Distance' indicates the distance covered during the event; 'High-Speed Distance' indicates the distance covered at speed $\geq$5.5 m/s during the event; and 'Unique Turning Angles' indicates the number of different directions of movement during the event. The average and standard deviation have been calculated for the statistical features derived from the speed and acceleration values of each action. In addition, the total count of accelerations and decelerations has been calculated for each event.

Although the Event dataset aggregates actions to represent higher-level contextual information, this aggregation reduces the temporal resolution and detail available in the raw feature set. To preserve vital granular information necessary for metric validation and detailed analysis, we retained the Action dataset as a parallel representation. This dual-dataset approach balances the need for event-level repre-

Table 3.12: Features extracted from each event in the Event dataset. Each event is described by temporal, spatial, speed, acceleration, and movement pattern features derived from the underlying sequence of actions.

Feature	Description
Start_Time	Timestamp or index where event begins
End_Time	Timestamp or index where event ends
Start_Second_from_Beginning	Event start time (in seconds) since the match (or half) began
End_Second_from_Beginning	Event end time (in seconds) since the match (or half) began
Num_Actions	Number of actions within the event
Action_Types	List of unique action categories occurring in the event
Action_Type_Counts	Count of each action type within the event (e.g., [3 Walking, 2 Running])
Action_Type_Perc	Percentage of each action type in the event (e.g., [0.6 Walking, 0.4 Running])
Action_Entropy	Shannon entropy calculated from action type distribution: $-\sum p_i \log(p_i)$
Action_Transitions	Number of type transitions (e.g., Walking→Running) within the event
Most_Common_Action	The most frequently occurring action type in the event
Event_Duration	Total event duration (seconds)
Mean_Action_Duration	Mean duration of actions within the event
Min_Action_Duration	Minimum duration among actions in the event
Max_Action_Duration	Maximum duration among actions in the event
Total_Distance	Total path length covered during the event (metres)
Avg_Speed	Mean speed during event (metres/second)
Max_Speed	Maximum speed during the event
Min_Speed	Minimum speed during the event
Speed_CV	Coefficient of variation (std/mean) of speed in event
Avg_Acceleration	Mean acceleration in event
Max_Acceleration	Maximum acceleration in event
Acc_CV	Coefficient of variation (std/mean) of acceleration
Speed_Trend	Slope of speed trend over event (via linear regression or difference)
Acc_Trend	Slope of acceleration trend over event
High_Speed_Distance	Distance covered above high-velocity threshold (metres)
Num_High_Speed_Actions	Number of "high-speed" actions (above a speed cutoff)
Num_Accelerations	Total number of acceleration bursts above threshold
Avg_Accl_X	Mean acceleration in the X axis
Avg_Accl_Y	Mean acceleration in the Y axis
Avg_Accl_Z	Mean acceleration in the Z axis
Turning_Angle_Sum	Sum of absolute turning angles across event (degrees or radians)
High_Acc_Events	Number of actions with acceleration above "high" threshold
Actions	Full list/sequence of raw actions during the event
Speed_Range	Difference between max and min speed in the event
Acc_Range	Difference between max and min acceleration in the event
Action_Type_Diversity	Count of different action types (cardinality) in the event
High_Intensity_Ratio	Fraction of actions/events tagged "high intensity" (e.g., running/sprinting actions)
Turning_Angle_Rate	Sum of turning angles divided by event duration
Acc_Variability	Standard deviation of acceleration during event
Speed_Variability	Standard deviation of speed during event

Table 3.13: Sample of the Event dataset obtained as a final step of the proposed multi-step framework. The names of some columns in the header have been shortened: player ID (P.ID), event ID (E.ID), start second (Start Sec.), end second (End Sec.), avg speed (Avg Sp.), std speed (Std Sp.), duration (Dur.), standard deviation duration (Std Dur.), distance (Dis.), high-speed distance (High-Speed Dist.), and unique turning angle (U.T.A.). Some of the existing columns have been omitted due to space reason.

P.ID	E.ID	Start Sec.	End Sec.	Start Lat	Start Lon	End Lat	End Lon	Avg Sp.	Std Sp.	Dur.	Std Dur.	Dis.	High-Speed Dist.	U. T. A.
152	4	23	27	54.55	−7.35	54.11	−7.35	5.3	0.3	4	0	23.5	3.1	2
152	5	27	29	54.51	−7.31	54.11	−7.35	3.2	0.1	2	0	5.9	0	1
152	6	29	39	54.21	−7.31	54.11	−7.31	5.7	1.7	10	2.6	60	6.7	3
152	7	29	43	54.01	−7.01	54.18	−7.39	3.3	1.2	14	3.4	36.9	0	4
152	8	43	48	54.11	−7.42	54.16	−7.30	6.3	0.5	5	1.9	41.2	11.1	1

sentation of data with the retention of fine-grained movement data, ensuring that no critical information is lost during feature engineering.

3.4 Case Study-Based Validation

This section presents two complementary case studies designed to demonstrate the effectiveness and versatility of the feature engineering framework. The first case study validates the Action dataset by comparing it against established sport science metrics and previous raw data while providing a profile of the actions. The second case study leverages the Event dataset to describe events and demonstrate its suitability for advanced predictive machine learning modelling. These case studies provide a comprehensive evaluation of the framework's ability to transform raw sensor data into actionable insights for both descriptive and predictive sports analytics.

3.4.1 Case Study 1: Analysis of Velocity Zones using Action Dataset

The motivation for this case study is to demonstrate that the engineered Action dataset is both valid (consistent with existing literature and original raw data) and capable of supporting deeper analysis of player activity.

Validation of Action Dataset Using Gaelic Football Research

To validate the effectiveness of the framework in engineering raw sensor data into discrete periods representing player movement within velocity zones, we compared running profiles obtained from the Action dataset with those published in prior Gaelic football research. This comparison confirms that our feature engineering preserves the integrity of raw sensor data, retains critical information despite aggregation, and produces movement profiles consistent with established Gaelic football research.

Three studies reporting elite Gaelic football match running profiles were selected

as comparators for this validation analysis: Malone et al. [23], Malone et al. [19], and Gamble et al. [54]. The selection of these studies was guided by four explicit criteria. First, all three studies examine elite male intercounty Gaelic football players, consistent with the population from which the present dataset was collected. Second, all three were published between 2016 and 2019, making them contemporaneous with the 2019–2021 data collection window and reflective of the physical demands of the game during the same years, prior to the more substantial rule changes that followed. Third, all three studies employed STATSports GNSS technology for data collection, the same manufacturer as the device used in the present work. Fourth, the three studies provide complementary dimensions of running profile coverage: Malone et al. [23] establishes absolute distance and speed benchmarks across positional roles; Malone et al. [19] characterizes the temporal decay of running output across quarters of the game; and Gamble et al. [54] reports relative distance per minute by velocity zone, which is the metric most directly analogous to the action-level output produced by the present framework. These three studies provide a multi-dimensional reference against which the validity of the framework can be assessed across absolute running volume, temporal patterning, and velocity zone distribution. A comparison was made between the profiles identified in these studies and those extracted from the proposed Action dataset. Only players who completed the full duration of the games were included in the analysis, except the goalkeeper.

Authors in [23] identified an average distance covered of 8,160.11 $\pm$ 1,482.02 m, with 1,731.29 $\pm$ 659.76 m covered at speed $>$ 4.72 m/s and 445 $\pm$ 169 m at speed $>$ 6.11 m/s performed by male players during elite Gaelic football games. The Action dataset produces similar values: the average distance is 8,633.8 $\pm$ 1,573.6 m, with 1,453.6 $\pm$ 552.7 m covered at speed $>$ 4.72 m/s and 503.5 $\pm$ 205.1 m at a sprinting speed ($>$ 6.11 m/s). Average and peak speed registered analogous results, with values equal to 1.8 $\pm$ 0.3 m/s and 8.4 $\pm$ 0.5 m/s in [23] and 1.9 $\pm$ 1.6 m/s and 8.4 $\pm$ 0.5 m/s in the Action dataset.

In [19], the authors observed a reduction in the distance covered between the first and other quarters of the game: −4.1% in the second, −5.9% in the third, and −3.8% in the fourth. In addition, they found that players reduced high-speed running distance (≥ 4.72 m/s) in the second (−8.8%), third (−15.9%), and fourth (−19.8%) quarters when compared to the first quarter. On the Action dataset we observed a similar reduction in the distance covered between the first and other quarters of the game: −6.6% in the second, −6.6% in the third, and −15.8% in the fourth. When comparing the high-speed running distance between the first and other quarters, the reductions were equal to −7.2% (second), −11.3% (third), and −17.1% (fourth).

Table 3.14 presents a direct comparison between relative distance covered per minute for different velocity zones in our dataset and reference values from [54]. Consistent with [54], only players who completed the entire duration of all games included in the study were considered in the analysis, except the goalkeeper. Walking, jogging, and running distances are well aligned with previous research. In particular, high-speed running and sprinting show an even closer match. This highlights the consistency of our results with established research, reflecting the robustness of our feature engineering process.

Table 3.14: Average (± standard deviation) distance covered (m/min) by player in each velocity zone in the Action dataset and in Gamble et al. (2019) [54].

Velocity Zone	**Action Dataset**	**Gamble et al. (2019)**
Standing (m/min)	0.1 ± 0.0	0.5 ± 0.1
Walking (m/min)	33.8 ± 5.7	29.8 ± 5.5
Jogging (m/min)	37.7 ± 10.4	35.1 ± 11.6
Running (m/min)	21.4 ± 7.8	18.8 ± 7.8
High Intensity Running (m/min)	8.6 ± 4.1	8.3 ± 3.4
Sprinting (m/min)	2.2 ± 1.9	1.7 ± 1.2

Despite marginal differences in mean values across velocity zones, the standard deviation for sprinting distance is notably higher in the Action dataset (2.2 ± 1.9 m/min) compared to Gamble et al. (1.7 ±1.2 m/min), representing a coefficient of variation of 86% versus 70% respectively. This elevated variance in the sprinting zone is consistent with the known effect of 1 Hz downsampling on high-intensity efforts:

averaging within-second speed values can introduce variability in the classification of brief sprinting actions near the zone boundary, as acknowledged in Section 3.2.1. Gamble et al. retained the full 10 Hz signal, which provides a more stable classification of short-duration high-speed efforts. Notwithstanding this difference at the sprinting zone, high similarity has been found between running profiles identified using the Action dataset and those reported in past research across all other velocity zones.

This demonstrates that, although the framework modifies the data structure and reduces granularity, it does not result in substantial data loss. Core information is preserved, enabling the extraction and definition of running profiles and activity levels from the Action dataset.

Validation of Action Dataset with Raw Sensor Data

Despite reducing the dataset size from over 76 million raw sensor rows to approximately 160 thousand rows describing actions (99.6% reduction), our analysis demonstrates that despite the reduction, the Action dataset preserves critical running performance information. Key player running metrics extracted from the Action dataset closely align with those derived from the raw sensor data, confirming minimal information loss.

For example, comparing total distances covered per player, the raw sensor data shows a range of 8,352 to 9,208 meters per game (Table 3.2 in Section 3.2.3), while the Action dataset averages 8,634 metres ($\pm$ 1,574 m). Similarly, distances covered at high speeds ($\geq$ 4.72 m/s) range from 1,402 to 1,744 m in raw data, with the Action dataset reporting a comparable mean of 1,454 m ($\pm$ 553 m). Also, distances at speed $\geq$ 6.11 m/s are stable, with raw data values between 441 and 556 m and the Action dataset averaging 504 m ($\pm$ 205 m).

Speed distribution characteristics remain preserved: the raw data mean speed is 1.97 m/s ($\pm$1.51 m/s) with a median of 1.50 m/s, while the Action dataset reflects a similar average speed profile (first half: 2.0 $\pm$ 1.5 m/s; second half: 1.94 $\pm$ 1.6

m/s), with patterns of slight median speed decrease in the second half consistent across both datasets.

These comparisons demonstrate that the feature engineering framework effectively aggregates raw sensor data into informative action segments. This approach maintains the integrity of essential performance metrics derived from the raw data, such as total distance covered, distances at high and sprinting speeds, and average and max speed.

These results demonstrate that the aggregation process preserves the running performance metrics of primary interest in this research (total distance, high-speed running distance, sprint distance, and average and peak speed) with negligible divergence from values derived directly from the raw sensor data. It should be acknowledged, however, that the aggregation is not without trade-offs. The reduction from over 76 million rows to approximately 160 thousand action-level rows necessarily involves the loss of within-second temporal granularity, the compression of fine-grained speed fluctuations within each action into summary statistics, and the discarding of the full 100 Hz accelerometer signal in favour of aggregated acceleration descriptors. The claim of information preservation is therefore specific to the running profile metrics validated here and should not be interpreted as a general assertion that no information is lost. Rather, the framework is designed so that the information most relevant to the analytical objectives of this research is retained, while the reduction in data volume substantially improves computational tractability and enables the action-level and event-level analyses that follow.

In-Depth Analysis of Action Dynamics by Game Segment and Velocity Zone

This section presents an exploratory analysis of the Action dataset, characterizing player movement through descriptive statistics, action counts, duration and distance profiles, velocity zone distributions, accelerometer summaries, and action transition probabilities. The purpose is to reveal patterns in player movement dynamics that

Table 3.15: Half by half analysis: action count, average, and maximum duration, sorted by count of actions.

Game ID	Half	Count	Duration Avg (s)	Duration Std (s)	Duration Max (s)
893	1	4,536	2.96	2.57	23
997	1	4,308	2.88	2.59	22
788	1	4,193	2.80	2.36	25
973	1	4,165	2.83	2.47	26
869	2	4,129	2.80	2.36	23
893	2	4,083	2.89	2.53	26
811	1	4,071	2.97	2.64	25
946	1	4,018	2.81	2.41	23
889	1	3,984	2.94	2.60	25
873	1	3,983	2.83	2.48	24
997	2	3,972	2.98	2.61	25
934	1	3,919	2.95	2.59	23
811	2	3,844	2.99	2.63	37
869	1	3,806	2.79	2.34	25
788	2	3,729	2.85	2.46	25
873	2	3,677	2.91	2.49	23
838	1	3,662	2.87	2.53	24
973	2	3,614	2.85	2.45	25
946	2	3,594	2.74	2.35	26
934	2	3,502	2.89	2.50	26
889	2	3,441	2.89	2.48	25
838	2	3,417	2.86	2.61	29

are made accessible by the action-level representation and that cannot be directly observed from fixed-window aggregate metrics or raw sensor data alone. Where comparisons are made between the first and second halves of games, paired-samples t-tests were applied to assess whether observed differences were statistically significant, with a significance level of $\alpha = 0.05$. First and second half values were paired within games to assess between-game differences. All other analyses in this section are descriptive in nature and are not intended as inferential comparisons with external sources.

The average speed during the first and second half is 2.0 ± 1.51 m/s and 1.94 ± 1.61 m/s, respectively.

Table 3.15 displays the number of actions and the average, standard deviation, and maximum duration for each half of each game. The overall average number of actions for each game is 3945 ± 286. The average duration of the actions is 2.9 ± 0.1 s. The average of the maximum duration of the actions is 25.2 ± 3s.

The resulting p-value for the t-test on the actions count is 0.004 (lower than the chosen level of significance, 0.05), showing a statistically significant difference for

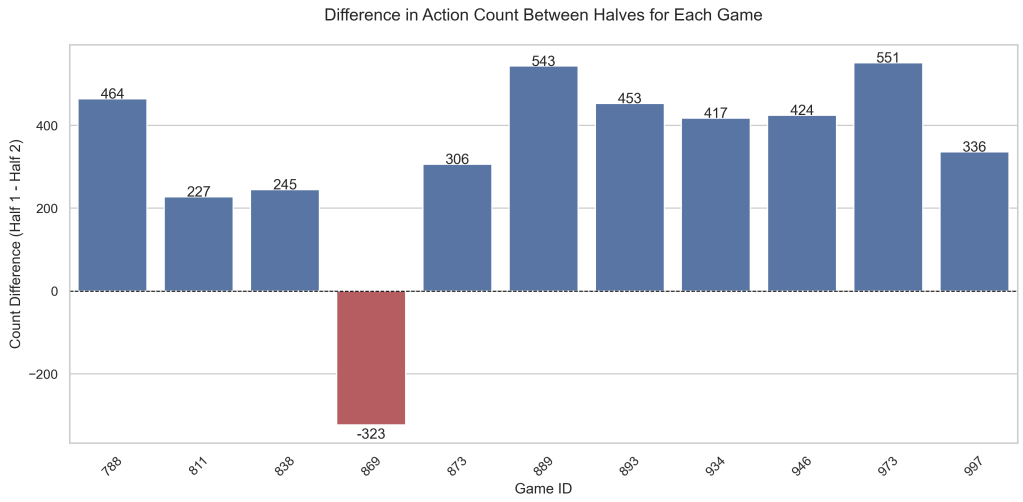


Figure 3.6: Difference in action count between first and second half per game.

the number of actions in the two halves of the game, which is almost always higher in the first half, except for game 869, as shown in Figure 3.6. This demonstrates that players tend to switch between velocity zones more often early in the game, while movements become steadier and involve fewer changes in the second half.

Table 3.16 shows an analysis of the labeled actions for each game. The average number of actions per game decreases as the speed increases: 'jogging' (4400.5 ± 226.5), 'running' (2402.5 ± 138.9), 'high-intensity running' (827.9 ± 54.3), and 'sprinting' (155.1 ± 14.6). For a given action, the average duration and distance are stable across different games. This suggests that players' behavior for each movement type is consistent regardless of the specific match.

A *t*-test identified a statistically significant difference in the average duration of 'high-intensity running' actions between the first half (2.2 ± 0.1 s) and the second half (2.0 ± 0.1 s). Regarding the other velocity zones, the average duration of 'jogging', 'running' and 'sprinting' actions are 3.4 ± 0.1, 2.3 ± 0.1, and 2.1 ± 0.2 s, respectively.

A statistically significant difference has been identified through a *t*-test for the average distance of 'high-intensity running' action between the first (13.1 ± 9.6 m) and second half (12.4 ± 8.7)

The remaining actions revealed similar outcomes when comparing the two halves.

Table 3.17 summarizes descriptive statistics for each velocity zone. The "Per-

Table 3.16: Count, average, maximum duration, and distance covered for each type of action during the sampled games, sorted by average distance.

Action	Game ID	Count	Duration Avg (s)	Duration Max (s)	Distance Avg (m)	Distance Max (m)
Sprinting	873	168	2.4	7	17.8	58.8
	973	149	2.3	9	17.1	71.1
	889	153	2.3	7	17.1	57.4
	869	178	2.2	8	16.5	60.8
	788	168	2.2	6	16.4	49.2
	838	131	2.1	8	16.1	62.4
	893	171	2.1	9	16.0	66.6
	997	146	2.0	6	15.5	47.4
	934	162	2.0	7	15.1	53.2
	811	136	2.0	7	15.0	52.5
	946	144	1.9	6	14.7	49.8
High-Intensity Running	869	911	2.2	11	13.6	68.2
	889	820	2.2	12	13.3	70.8
	997	868	2.2	13	13.2	79.3
	811	784	2.1	11	13.1	71.5
	973	837	2.1	14	13.00	86.8
	893	906	2.1	10	12.8	65.0
	838	726	2.0	10	12.6	63.0
	873	836	2.0	10	12.5	63.0
	946	780	2.0	10	12.4	60.0
	934	781	2.0	10	12.3	64.0
	788	858	2.0	11	12.1	67.1
Running	997	2589	2.3	21	10.9	96.6
	893	2,653	2.4	20	10.9	94.0
	811	2,417	2.3	24	10.8	112.8
	973	2,449	2.3	18	10.8	86.4
	788	2,451	2.3	22	10.6	110.0
	869	2,468	2.3	20	10.6	96.0
	873	2,337	2.3	16	10.6	76.8
	889	2,313	2.3	16	10.6	76.8
	934	2,256	2.2	16	10.4	76.8
	946	2,357	2.2	15	10.3	69.0
	838	2,138	2.2	14	10.1	61.6
Jogging	811	4578	3.5	37	10.5	107.3
	934	4,222	3.5	26	10.4	72.8
	893	4,889	3.4	26	10.3	85.8
	889	4,139	3.4	25	10.3	82.5
	997	4,677	3.4	25	10.3	77.5
	838	4,084	3.4	29	10.1	84.1
	873	4,319	3.3	24	10.0	74.4
	973	4,344	3.3	26	9.9	80.6
	788	4,445	3.3	25	9.8	80.0
	946	4,331	3.2	26	9.7	84.0
	869	4,378	3.2	25	9.6	70.0

Table 3.17: Summary statistics for each velocity zone.

Action	Percentage	Duration	Distance	Avg Speed	Std Speed	Avg Acc.	Avg Accl X	Avg Accl Y	Avg Accl Z
Standing	8.3	3.9	0.1	0.04	0.03	0.40	-0.019	0.591	0.357
Walking	31.9	8.7	10.5	1.23	0.32	1.30	-0.014	0.639	0.357
Jogging	35.7	3.4	10.2	2.94	0.49	2.01	-0.006	0.629	0.376
Running	17.3	2.5	11.5	4.60	0.33	2.06	-0.003	0.616	0.376
High Intensity Running	5.8	2.2	13.9	6.05	0.32	2.08	0.003	0.613	0.365
Sprinting	1.0	2.3	17.5	7.37	0.23	2.08	0.014	0.632	0.332

centage" column indicates the proportion of each action relative to the total number of detected actions, revealing that jogging is by far the most prevalent activity, accounting for nearly 35.7% of all occurrences, followed by walking (31.9%) and running (17.3%). This distribution reflects the typical activity profile in field sports, where moderate-intensity movement represents the vast majority of movements and high-intensity efforts are comparatively rare.

Distance per action increases steadily, with sprinting actions characterized by the greatest distance covered per action (17.5 m), despite having a lower duration. Average acceleration also rises modestly with action intensity. The mean values for Accl X remain close to zero, ranging from almost negative during jogging and running to slightly positive during sprinting. These near-zero values suggest that, for each action, forward and backward accelerations are largely balanced. The average Accl Y values are positive and relatively consistent across all actions, with only minor variation. This indicates a small but steady lateral acceleration component, which may reflect directional adjustments made during movement. The similarity in Accl Y across actions suggests that, regardless of overall intensity, the lateral movement demands remain relatively constant. For Accl Z, the average values are positive but modest, and they show a slight decrease as action intensity increases. This vertical acceleration component is associated with the up-and-down motion inherent in running and sprinting. The small reduction in Accl Z for higher-intensity actions may reflect a more horizontal, less vertically oscillation at maximal speeds. This can be explained by the tendency of players to prone forward during high-intensity actions to minimize unnecessary vertical movement and maximize speed.

The action transition probability matrix in Figure 3.7 defines how likely it is for a player to move from one activity (on the y-axis) to another (on the x-axis). As ex-

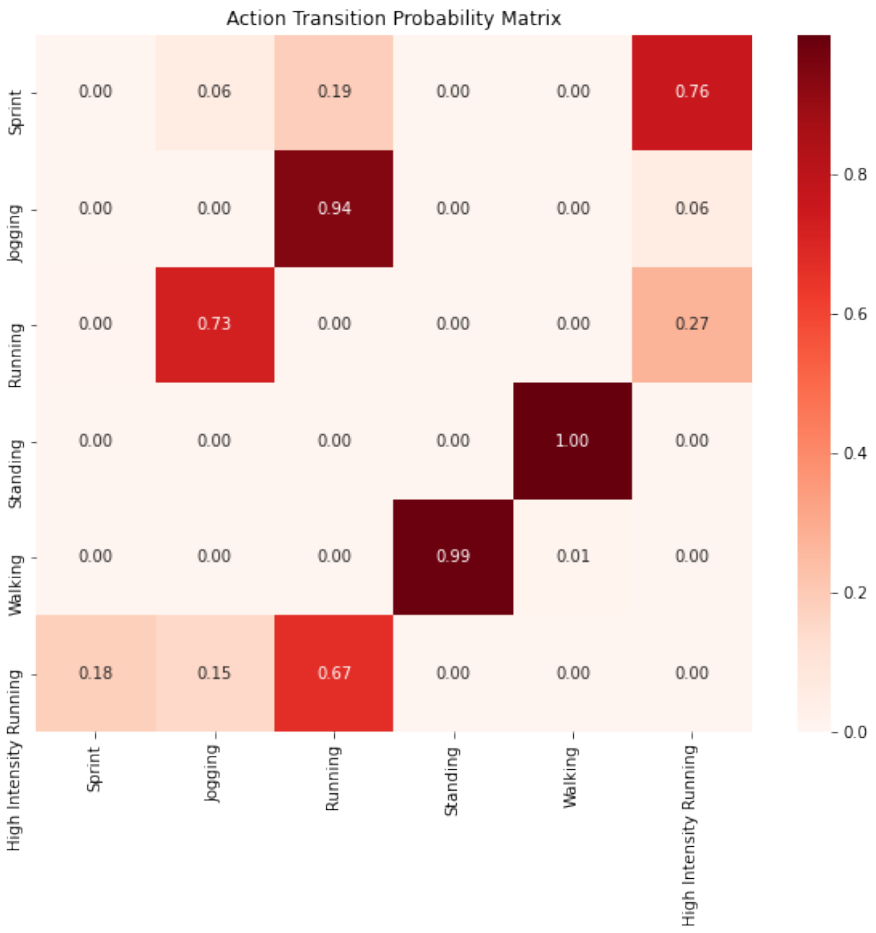


Figure 3.7: Probability matrix of action transitions. The starting action is on the y-axis, while the following action is on the x-axis. Each value in the matrix represents the probability of going from the action on the y-axis to the action on the x-axis.

pected, the vast majority of actions are followed by the next or the previous velocity zone. For example, running actions only moved up to high-intensity running (27%) or down to jogging (73%). Interestingly, for some actions, there is a probability to lower down to two or more velocity zones, such as sprinting to jogging (0.06%), sprinting to running (0.19%), and high-intensity running to jogging (0.15%).

3.4.2 Case Study 2: Analysis of Events and Predicting Distance at High Speed using Event Dataset

This section aims to validate the Event dataset and to present a predictive experiment using data from the Event dataset. Each row of the Event dataset describes an event performed by a player during a game. As was done with the Action dataset,

the Event dataset is validated by comparing key metrics with those reported in previous research. For each player who completed the full duration of the games (except the goalkeeper), the average distance, average speed, and average max speed were computed. Minimal differences were found with the Action dataset: the average distance was identical, while average and peak speeds differed by only 0.2 and 0.3 m/min, respectively. Other metrics could not be directly compared due to the aggregation of velocity zones into events. However, given the similarity in these fundamental features between the event and Action datasets, it is possible to validate the Event dataset for these specific metrics. This confirms that the aggregation process maintained data integrity and did not lose information or generate spurious outputs.

Descriptive Analysis of Event-Level Activity

In this section, event counts, durations, distances, and high-speed activities were analyzed to demonstrate how features extracted from events capture the intensity, frequency, and distribution of active periods between resting states during match play in Gaelic football. To focus on active events, those events composed only of standing and/or walking actions were excluded from the analysis.

After excluding events consisting solely of standing and/or walking actions, the dataset was reduced to 3,5389 active events The average number of events during the first and second halves is 1,657.3 ± 106.7 and 1,538.2 ± 94.9, respectively. The highest number of events during the first half is 1,937, with 7.1 ± 6.1 s of average duration and 63 s of maximum duration for an event (Table 3.18). The highest value for the second half is 1,669, with an average duration of 7.1 ± 6.2 seconds and a maximum duration of 72 seconds for an event.

The average number of actions per event is 2.55 ± 2.25, with 1 as the minimum and 21 as the maximum. This indicates that each event consists, on average, of between two and three discrete actions, suggesting that events generally capture short sequences of player activity rather than extended or highly complex sets of

Table 3.18: For each game-half, the table reports the total count of detected events, average and maximum event duration (seconds), average and maximum event distance (meters), and average/maximum distance covered above the high-speed running threshold.

Game ID	Half	Events (c)	Dur. Avg (s)	Dur. Max (s)	Dist. Avg (m)	Dist. Max (m)	H. S. Dist. Avg (m)	H. S. Dist. Max (m)
893	1	1,947	7.0	48	25.4	189.5	3.3	117.4
893	2	1,604	7.5	54	28.1	214.7	4.8	113.5
934	1	1,554	7.5	49	27.6	209.3	4.1	92.7
934	2	1,428	7.2	44	26.5	163.5	4.0	70.6
997	1	1,756	7.2	52	26.8	236.8	4.3	87.0
997	2	1,645	7.3	48	26.9	191.4	3.7	108.8
973	1	1,661	7.2	45	26.7	198.8	4.0	98.1
973	2	1,567	6.7	48	25.3	226.2	4.3	108.6
838	1	1,568	6.8	48	24.8	188.9	3.5	82.4
838	2	1,450	6.8	38	25.1	156.9	3.9	91.5
946	1	1,668	6.9	38	25.2	148.2	3.8	86.9
946	2	1,482	6.8	45	24.8	179.0	3.7	93.5
889	1	1,631	7.3	40	26.8	182.9	4.4	99.9
889	2	1,369	7.4	72	27.7	264.1	4.9	126.2
811	1	1,708	7.2	63	26.4	216.7	3.7	86.5
811	2	1,637	7.2	52	26.0	190.4	3.6	80.2
869	1	1,575	6.9	49	25.6	188.8	4.4	112.9
869	2	1,676	7.0	38	26.5	158.9	5.1	97.7
788	1	1,654	7.3	44	27.0	199.0	4.1	89.0
788	2	1,594	6.8	52	24.9	236.9	4.0	84.8
873	1	1,590	7.2	38	26.8	182.7	4.6	82.5
873	2	1,559	7.0	49	25.7	190.2	3.8	82.5

actions. The relatively high standard deviation (2.25) means that the vast majority of events are composed of a few actions. This proves that the vast majority of player movement transitions from rest to activity and back to rest every few seconds, reflecting the intermittent nature of the sport.

Table 3.18 reveals consistent average event durations (around 7 seconds) and event distances (about 25–28 m) across games and halves, indicating stable patterns of play despite some variability in maximum event lengths and distances.

Figure 3.8 demonstrates that the event count is generally higher in the first half than in the second half, with median values around 1,650 for the first half and about 1,570 for the second half. For the first half, the interquartile range is relatively compact, indicating that most games have a similar event count, except for one outlier. The box delimited by the interquartile for the second half is visibly wider, showing greater variability and reduced numerosity in event counts across games.

The frequency distributions of events duration and distance in Figures 3.9, 3.11, and 3.10 are right-skewed, with the vast majority of actions being short in duration and low in distance, consistent with the intermittent nature of Gaelic football. Occasional peaks at longer durations (for example around 37 and 48 seconds visible in

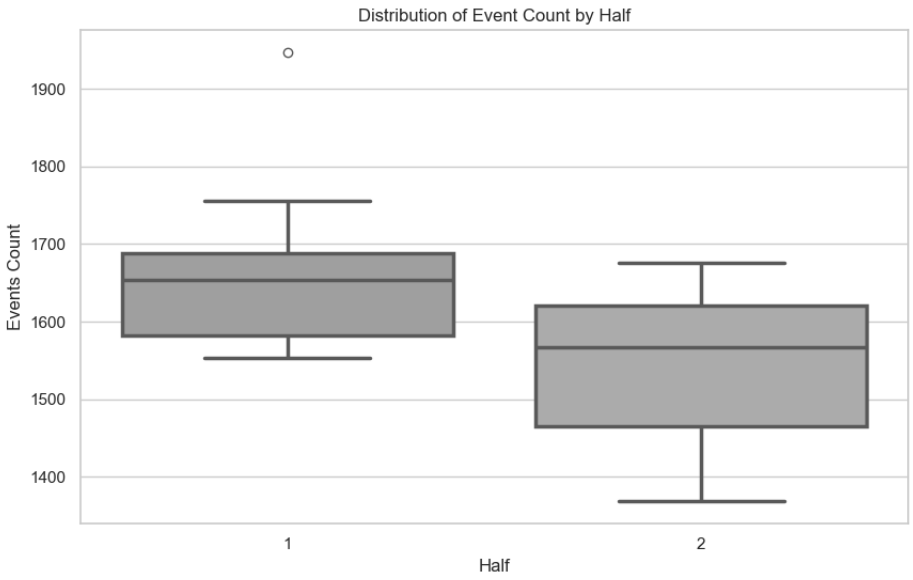


Figure 3.8: Boxplot of the distribution of event count by half.

the tail of the duration distribution) are likely attributable to game-specific interruptions such as frees, line-ball stoppages, or injury breaks, during which a player may sustain a low-speed action for an extended period without a velocity zone transition triggering a new action boundary. It is also possible that in a small number of cases, brief GNSS signal degradation may have prevented zone transitions from being detected, artificially extending an action. These occurrences are rare and do not affect the overall action-level representation, as evidenced by the stable average durations reported across games in Table 3.15.

High-speed events are a key aspect of the game, and as a result, predicting them is the central objective of the following analysis. In addition, knowing in advance the amount of high-speed distance covered by players can be beneficial for sports coaches to manage and optimize players' time on the pitch, to avoid fatigue or prevent injuries.

Predicting High-Speed Running Events

An event-level dataset was constructed to characterize the activity levels of players during matches. To demonstrate the potential of this dataset for machine learning

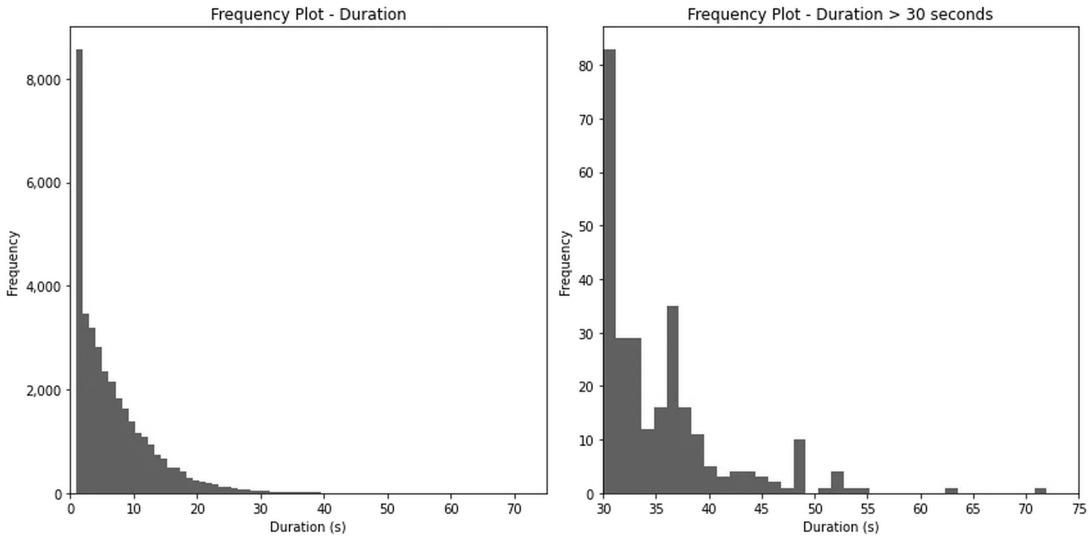


Figure 3.9: **Left**: Frequency of events per duration. **Right**: Frequency of events per duration $\geq$ 30 s.

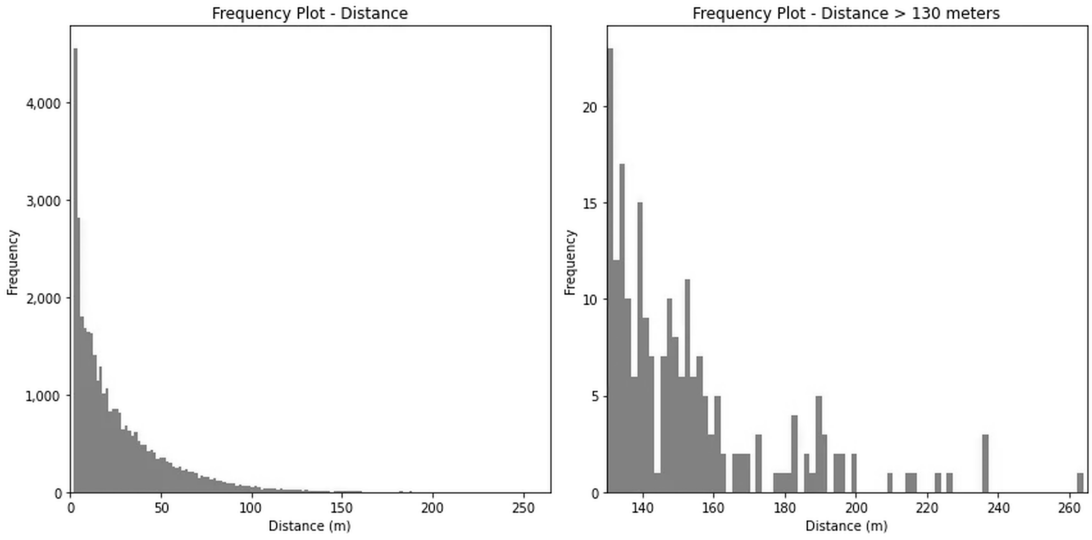


Figure 3.10: **Left**: Frequency of events per distance. **Right**: Frequency of events per distance $\geq$ 130 m.

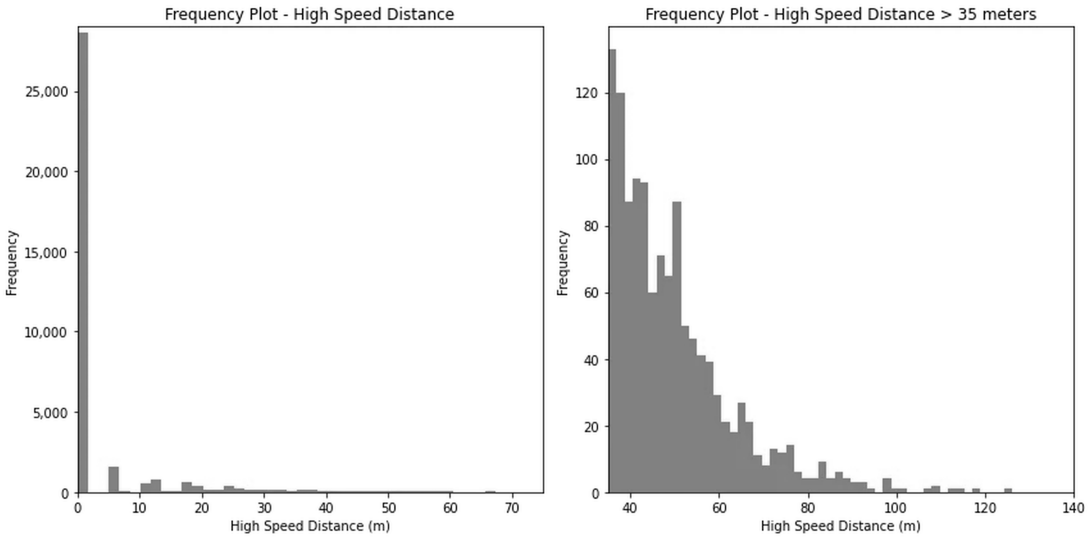


Figure 3.11: Frequency of events per high-speed distance. **Right**: Frequency of events per high-speed distance $\geq$ 35 m.

applications, an experiment was designed to predict high-speed distance (speed > 5.5 m/s) during the second half of a game.

This predictive task is enabled by first pre-processing and reshaping the Event dataset through relevant feature engineering and target variable preparation. The new manipulation of the Event dataset represents the input dataset for the machine learning models. In the following steps, we detail the data processing pipeline.

- **Input Data Definition**. The Event dataset, obtained at the end of the multi-step framework and shown in Table 3.13, represents the input dataset. This specific task utilized only the following columns: GameID, Half, PlayerID, and High-Speed Distance, excluding all the others.

- **Testing Games Using Leave-1 Game Out- Cross Validation**. For each player, the second half of one game was used as the test set, and the first half plus the remaining games as the training set. This method is called Leave-1 Game Out- (LOGO) Cross Validation, and it permits assessing the model's prediction accuracy on multiple datasets (1 per each second half of the available games) while replicating a real-case scenario (using available games and the first half to predict high-speed distance in the second half). The machine learning model was trained on the training set, and results were collected on

the test set. Then, the experiment was repeated by changing at each time the tested game and the training set.

- **Preparing Event Data for High-Speed Distance Forecasting**. After selecting one game as the test set, the Event dataset is processed to predict players' high-speed distance based on data from the first half and other games. Predicting at the event level allows the model to follow how a player's effort changes over time during the half, which is something a single total prediction cannot show. In addition, forecasting a cumulative total distance would require a high amount of player-game samples, while predicting event-level sequences uses more data points (one prediction per event instead of one per game) from each game. This means the model can learn well even with fewer overall games because it uses many smaller pieces of data from each game. Instead of needing many whole-game examples, the model benefits from the increased number of event-level data points, which speeds up learning and improves accuracy. This approach also allows the model to better learn how a player's effort changes over time. The steps to process the Event dataset are the following:

  - From the Event dataset, the second half of the selected game was extracted. The dataset was then filtered to use only players who completed the full game, discarding the rest.
  - For players who played the entire game, the time series data of high-speed distance from the remaining games, plus the first half of the tested game, were merged into a single time series. The first half of the tested game was placed at the end of this time series, as shown in Figure 3.12.
  - Then, for each player, a sequence of 30 consecutive high-speed distance values was used as a new set of features, and the variable to predict was the high-speed distance at the 31st position (Table 3.19). In summary, 30 consecutive rows in the Event dataset were used as columns in the new dataset to predict the 31st row. The dataset displayed in Table 3.19

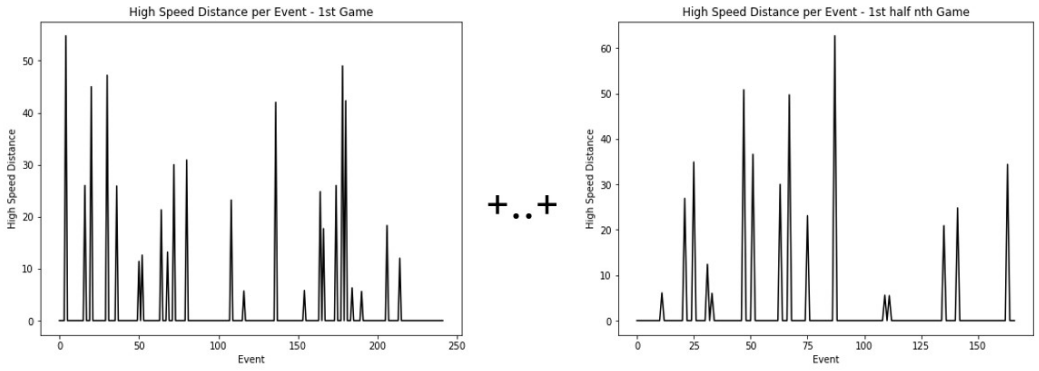


Figure 3.12: The time series of high-speed distance for the remaining games are merged to form a unique time series. The first half of the tested game is placed in the last position. These data constitute the training set for the predictive experiments.

Table 3.19: Structure of the transformed dataset used for the prediction of high-speed distance (H.S.D.) covered by players during the second half of a game. The **target feature** to predict is the variable 31^{st} **H.S.D** located at the end of the sequence.

P.ID	**1st H.S.D.**	**2nd H.S.D.**	**3rd H.S.D.**	**...**	**...**	**...**	**29th H.S.D.**	**30th H.S.D.**	**31st H.S.D.**
152	0	0	0	..	..	..	3.2	0	5.1
152	0	0	0	..	..	..	0	5.1	0
152	0	0	15.4	..	..	..	5.1	0	0
152	0	15.4	0	..	..	..	0	0	21.3
152	15.4	0	2.6	..	..	..	0	21.3	0

represents the training test, where the variable to predict is the value in the **31st H.S.D.** column.

– While the learning phase uses data from prior games together with the first half of the target game, during the testing phase, the test set is formed by the model's predictions at each iteration. This approach is known as recursive forecasting. In this method, the model is trained on historical data, and rather than having a separate test set, it uses its predictions as inputs for future predictions. Assuming N the number of events performed by player i during the first half of the tested game, the realization of the testing set can be described as follows:

1. The last 30 high-speed distance values of the first half are used as input (test set) by the model to generate the first prediction.
2. The last 29 high-speed distance values of the first half and the first prediction of the second half are used as input by the model to gen-

erate the second prediction.

3. Finally, the Nprediction uses the previous N-1 predictions. If $N-1 < 30$, the remaining inputs are taken from the final high-speed values in the first half.

At the start of the second half, the number of player events for the half is unknown. For the purpose of the experiment, this number is assumed to be equal to the number of events in the first half. Thus, the recursive prediction process is repeated as many times as the number of events performed by the player during the first half. Finally, the predicted distances are summed to obtain the final high-speed distance covered by the player.

- **Testing Machine Learning Models for Player-Specific Prediction.** Table 3.19, obtained in the previous step, represents the input dataset for the machine learning models. Three distinctive machine learning models were tested to predict their effectiveness in predicting high-speed running distance. Instead of using a single model with data from all players, the same three machine learning models were trained and tested separately for each player. This approach forces the model to learn patterns specific to each player, recognizing that different players may have different behaviors. The set of features (30 consecutive values of distance at high speed) used is the same for each model (and each player). The machine learning models implemented to solve the predictive task were the following:

  - XGBoost [102] with the following hyperparameters: *n_estimators*, 100; *learning_rate*, 0.1; *max_depth*, 3; *min_child_weight*, 1; *colsample_bytree*, 1; *subsample*, 1; *reg_alpha*, 0; *reg_lambda*, 1.
  - A Long Short-Term Memory Network (LSTM) formed by the following layers: Bidirectional LSTM (unites = 64, activation = 'tanh') [103], Bidirectional LSTM (unites = 32, activation = 'tanh'), Dropout (0.20),

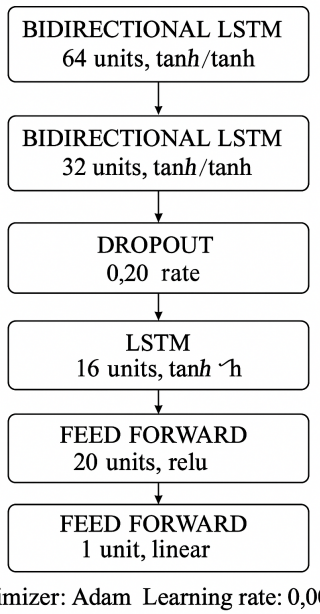


Figure 3.13: Architecture of the LSTM-based neural network proposed.

LSTM (unites = 16, activation = 'tanh'), Feed Forward (units = 20, activation = 'relu'), and Feed Forward (units = 1, activation = 'linear'). The optimizer used was 'Adam' [104], while the learning rate and the number of epochs were 0.001 and 50, respectively. Figure 3.13 presents the architecture of the LSTM neural network, highlighting the configuration of each layer and its specific parameters.

– A Convolutional Neural Network (CNN) formed by the following layers: one-dimensional convolutional (filters = 64, kernel_size = 5, activation = 'relu') [105], one-dimensional convolutional (filters = 32, kernel_size = 3, activation = 'relu'), one-dimensional convolutional (filters = 10, kernel_size = 5, activation = 'relu'), one-dimensional MaxPooling (pool_size = 2) [106], Flatten, Feed Forward (40, activation = 'relu'), Dropout (0.2), and Feed Forward (1, activation = 'linear'). The optimizer used is 'Adam', while the learning rate and the number of epochs were equal to 0.001 and 50, respectively. The structure of the CNN network is shown in Figure 3.14.

Both LSTM and CNN networks were implemented in Keras [107].

For each player, the number of events in the first half is assumed for the

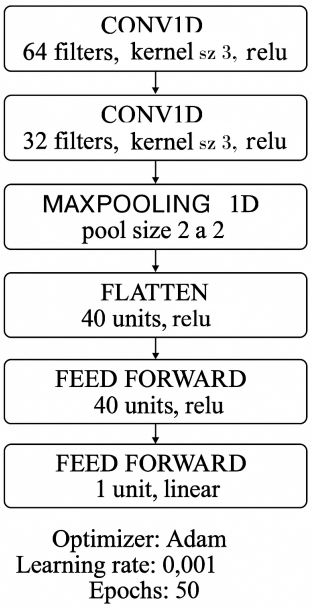


Figure 3.14: Architecture of the CNN-based neural network proposed.

second half, with high-speed distance predicted for each event. These predicted distances are summed to obtain the total predicted high-speed distance for the second half.

- **Predictive Experiment Results**. Table 3.20 presents the average Mean Absolute Error (MAE) for XGBoost, LSTM, and CNN when predicting total high-speed distance across a series of competitive games. The mean absolute error (MAE) measures the average absolute difference between the predicted values and the actual target values. MAE was used in this study because it is less sensitive to outliers than other metrics. In formula:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

where y_i are the actual target values, $\hat{y}_i$ are the predicted values, and N is the number of observations.

Here, calculating the mean absolute error (MAE) for each event was not possible because the number of events in the second half is unknown at halftime.

For this experiment, calculating MAE scores for each specific event was not

feasible due to the assumption of an equal number of events between the first and second halves. However, since the actual number of events performed by a player in the upcoming half was unknown before its start, it was decided to predict, for each player, the high-speed distance per event based on a number of events equal to those performed by the same player during the first half of the same game. This approach assumes that the number of events does not differ substantially between halves for players who completed both. The absolute error between the total predicted and actual high-speed distances across these events was computed for each player, and the mean of these per-player errors within a game represents the reported MAE for that game.

Across all games, XGBoost achieved the lowest overall average MAE at approximately 65 m, followed by LSTM at 87 m and CNN at 103 m. However, per-game errors for XGBoost range from 44.5 m to 106.9 m, and in individual player predictions errors exceeding 200 m were observed, representing a substantial proportion of the total high-speed distance a player might cover in a half. Model performance varies considerably from match to match, which may reflect differences in game context, tactical intensity, player fitness, and the limited number of games available for training under the LOGO cross-validation scheme. These results should therefore be interpreted with caution. While XGBoost consistently outperformed LSTM and CNN across all tested games, the magnitude of errors in some cases limits the practical utility of the current model for precision workload management. The findings are best understood as a proof-of-concept demonstration that the Event dataset structure is compatible with sequence-based machine learning tasks, rather than as a fully validated predictive tool ready for applied deployment.

To contextualize these results, the MAE can be expressed relative to the mean high-speed distance per player per half. Based on a full-game mean high-speed distance of 1,453.6 m, the estimated mean per half is approximately 727 m, giving an average relative MAE of approximately 9.5% for XGBoost, with

per-game values ranging from 6.1% to 14.7%. No published studies predicting high-speed distance at event level in Gaelic football or in sufficiently analogous contexts were identified, and direct benchmarking against other models therefore remains a direction for future work.

Several limitations of this predictive experiment should be acknowledged. First, the dataset comprises only 11 games from a single team, which restricts the volume of training data available per player and limits the generalisability of the findings. Second, the assumption that the number of events in the second half equals the number in the first half introduces a structural approximation that may propagate error into the final summed prediction. Third, no contextual features (for example, match score, opposition quality, or positional role) were included in the model, all of which are known to influence high-speed running output. Future work with larger multi-team datasets, richer feature sets, and systematic hyperparameter optimization would be necessary to establish whether the predictive performance demonstrated here generalizes beyond this sample.

Table 3.20: Average Mean Absolute Error (MAE) per model for each tested game. The MAE is computed on the total high-speed distance predicted per player.

Game ID	XGBoost	LSTM	CNN
788	**83.9**	110.2	140.3
811	**65.6**	85.1	90.8
869	**106.9**	130.4	180.3
893	**82.9**	105.3	120.6
934	**60.4**	75.3	88.5
838	**56.6**	70.4	85.2
873	**70.0**	85.9	93.0
889	**62.5**	78.4	85.4
973	**60.8**	75.2	95.1
997	**67.8**	90.4	99.2
946	**44.5**	57.8	65.3

To illustrate the model's performance at the individual player level, Figure 3.15 presents the true and XGBoost-predicted total high-speed distances for each player in Game 788. The predictions demonstrate variation in accuracy across

players, with some predictions closely matching actual performance while others show larger deviations. For example, Player 146's predicted workload for the second half is 576.9 m, very similar to previous performances such as 603.2 m in Game 946 and 595.8 m in Game 838, but higher than the 423.9 m actually covered in the current match. Player 152's true high-speed distance in the second half of Game 788 is 359.6 m, substantially lower than the predicted 517.2 m, and less than a previous peak value of 765.7 m (Game 811). Player 163 ran 428.9 m at high distance, with 582.5 m predicted. This overestimation contrasts with prior performances in the 400-500 m range across other games.

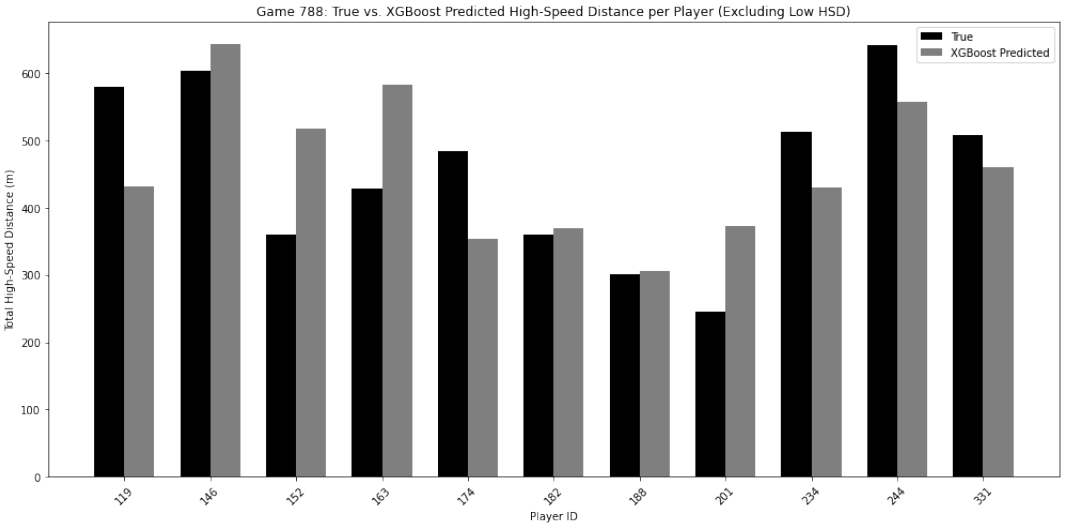


Figure 3.15: Comparison of true and XGBoost predicted total high-speed distances for each player in Game 788.

Figure 3.16 presents a detailed comparison of Player 152's true and XGBoost-predicted total high-speed distances across multiple games, demonstrating how predictions vary both within and across matches. In the vast majority of games, predicted and true values are very close. However, for some games, this difference is high. In detail, in Game 788, the model predicts a significantly higher workload (517.2 m) compared to the true recorded distance (359.6 m). On the contrary, in Game 889, the model prediction is greatly lower than the actual workload (573.4 m predicted vs. 815.1 m true). These prediction errors, while substantial, fall within the range of the model's average MAE of approximately 65 m when considered across all predictions.

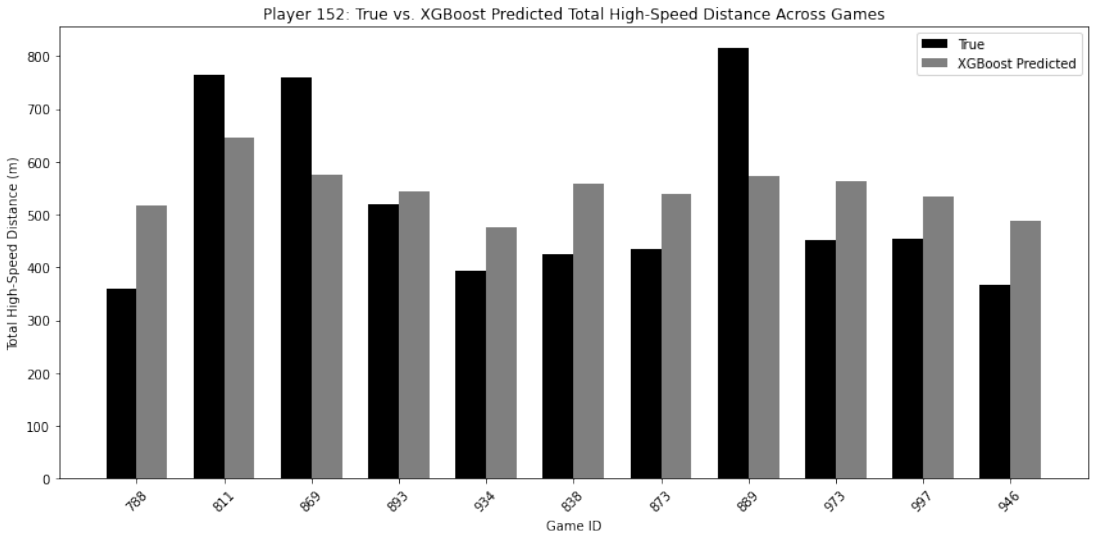


Figure 3.16: Comparison of Player 152's true and XGBoost predicted total high-speed distances across games.

3.5 Summary

Despite advances in wearable sensor technology, there remains a gap between usable knowledge (player movement and workload) and raw sensor data. This chapter presented a multi-step feature engineering framework that transforms the high-frequency, sequential sensor data into the Action and Event datasets. The Action dataset segments continuous player movement into discrete velocity-based actions, preserving information about timing and type of activity. Instead, the Event dataset aggregates these actions into periods of activity separated by rest states. The resulting datasets not only align with running profiles in Gaelic football but also offer the possibility of advanced analytics and predictive tasks.

This approach maintains essential player information while reducing data size (99.8%) from original raw data, which significantly speeds up data processing and analysis without losing critical details.

The practical relevance of the framework was further demonstrated through predictive modelling experiments using the Event dataset to forecast the high-speed running distances in the second half of games for individual players. Among the machine learning models tested, XGBoost achieved the lowest prediction errors across all games, with an average MAE of approximately 65 m per player per half. However,

errors varied substantially across games and individual players, with some predictions exceeding 200 m. These results serve as a proof-of-concept demonstration of the Event dataset's compatibility with sequence-based predictive modelling, rather than a fully validated applied tool, and future work with larger datasets will be necessary to establish the generalizability of these findings.

In addition, the value of this framework lies in making the sequential structure of player movement visible and analyzable. By segmenting continuous sensor data into discrete actions and grouping them into events, practitioners can examine how movement sequences are composed. Identifying the most common sequences and their associated durations and intensities allows coaches and conditioning staff to design training drills that replicate the actual movement patterns players experience during competitive match play.

Chapter 4

Interpreting Player Movement Using a Spatial Grid Model

The work presented in this chapter has been published in:

- Antonini Valerio, Michael Scriney, Alessandra Mileo, and Mark Roantree. "Using a Spatial Grid Model to Interpret Players Movement in Field Sports." In *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 395–400. Cham: Springer Nature Switzerland, 2024.

4.1 Introduction

Chapter 3 presented the development and validation of a framework that converts raw high-frequency sensor data into structured datasets. This enables detailed characterization of player running activity, assessment of movement intensities across velocity zones, and the use of predictive models for high-speed running distance. This work established a clear, reproducible framework for analyzing individual player activity and transitions between intensity levels.

However, while these datasets effectively characterize when and how intensely players move, they represent the spatial dimension only as individual coordinates.

This characteristic is shared with most sensor-based performance analyses in the existing sport science literature, where the movements at different speeds in the playing area are rarely examined [46], [108].

Addressing this gap is important because it allows practitioners to understand key aspects of match demands. Specifically, the research question (RQ2) investigates whether graph-based representations can reveal the following:

- **Critical transition zones:** Which zones of the playing area frequently serve as important transition points where players move between different areas, identified through betweenness centrality analysis.
- **Spatial communities with distinct workload profiles:** How groups of spatial cells form natural communities representing areas of coordinated player activity (identified through community detection algorithms) with distinct running intensity profiles.
- **Temporal evolution:** How the transition zones, spatial communities, and their associated running profiles change across consecutive time windows.

The Time-Window Spatial Activity Graph framework that will be introduced in this chapter leverages the Action dataset, realized in the previous chapter, to construct a series of spatial movement networks, capturing how player movements are organized through the playing areas. It provides a novel spatial perspective on player behavior that addresses a gap identified in the broader team sports literature: while GNSS-based analyses have extensively characterized the temporal and intensity dimensions of player movement [46], [23], the spatial organization of these movements across the playing area has received considerably less attention, with spatiotemporal analysis in team sports having concentrated predominantly on football and basketball rather than other invasion sports such as Gaelic football [108].

The novel contribution of this chapter is the TWG framework itself: the integration of the Action dataset with spatial tessellation, the rolling time window

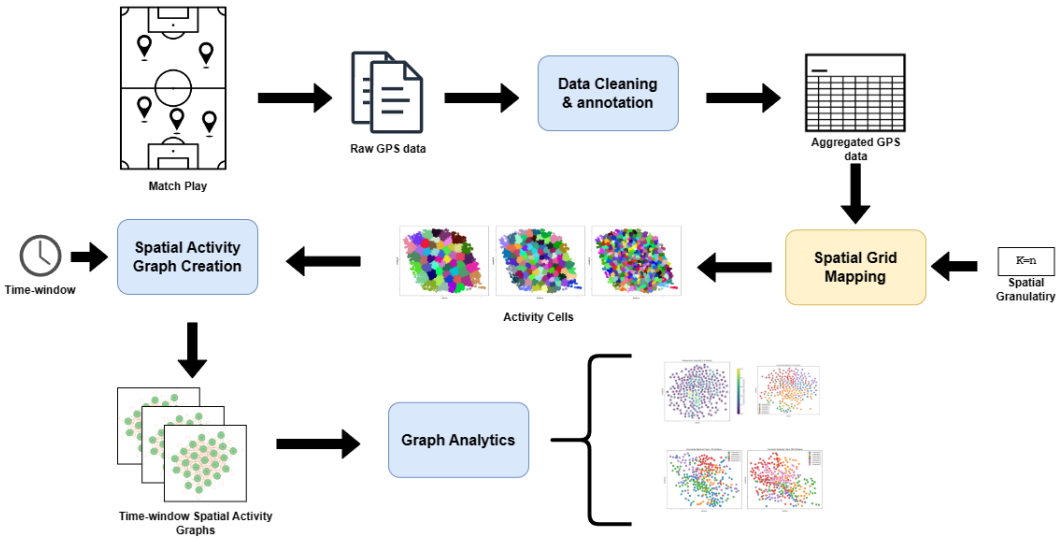


Figure 4.1: Diagram of steps involved in realizing the Time-Window Spatial Activity Graph.

construction of spatio-temporal multi-graphs, and the edge weighting scheme ($w = d \cdot \bar{Speed}^{1.5}$), applied to wearable sensor data in Gaelic football performance analysis. The analytical components used within the framework, specifically betweenness centrality [109], Louvain community detection [110], and the tessellation algorithm [111], are established methods from the literature. The novelty lies in their integration into a unified framework and their application to this specific domain.

4.2 Methodology

This section presents a graph-based framework that converts the Action dataset into a spatio-temporal graph, enabling analyses achievable only with this specific data representation. The steps to realize the framework are summarized in Figure 4.1. The objective of this section is to expose the required methodology to represent the actions of Gaelic football players of a team during a game as a graph, to extract patterns and insights about players' movement dynamics in the playing area, otherwise difficult or impossible to obtain from any other type of data representation.

Each step of the framework is illustrated using a running example from Game 934. This example helps demonstrate spatial cell mapping, graph construction, and graph analytic results.

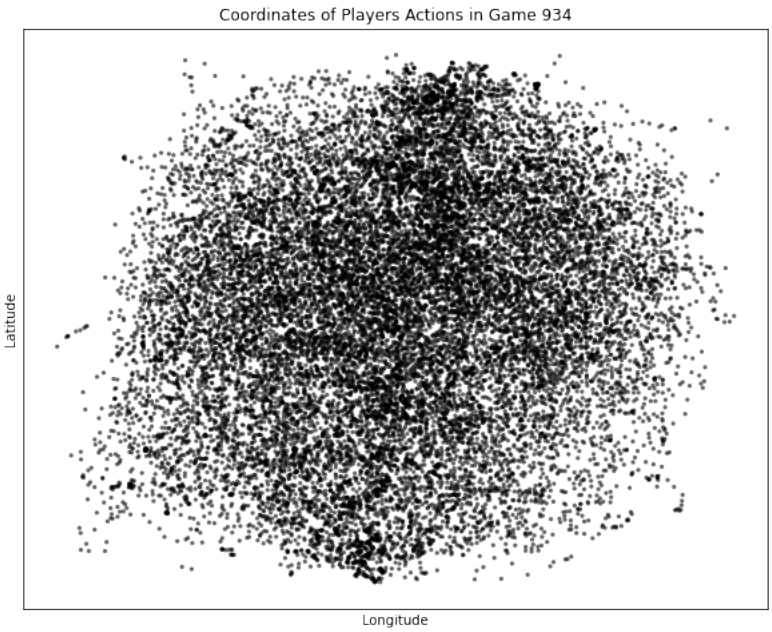


Figure 4.2: Coordinates of players' actions in Game 934.

The graph-based framework is composed of the following steps.

4.2.1 Input Data

The input data of the framework is represented by the Action dataset described in Chapter 3.3. The Action dataset describes the journey of players during games as a set of consecutive actions, expressed as consecutive seconds spent in the same velocity zone with a starting and ending location given by a pair of latitude and longitude coordinates. Figure 4.2 shows the coordinates of player actions during Game 934, where each coordinate represents the starting or ending point of each action during the game.

4.2.2 Spatial Cells Mapping

The next step is the segmentation of the playing area into cells. Dividing a playing area into non-overlapping, square cells using the GNSS coordinates covered by players during the game is subject to the orientation of the area on the globe, which makes this process not scalable and automatic. In addition, not every portion of the playing area is covered during a game. To analyze player movements, the playing

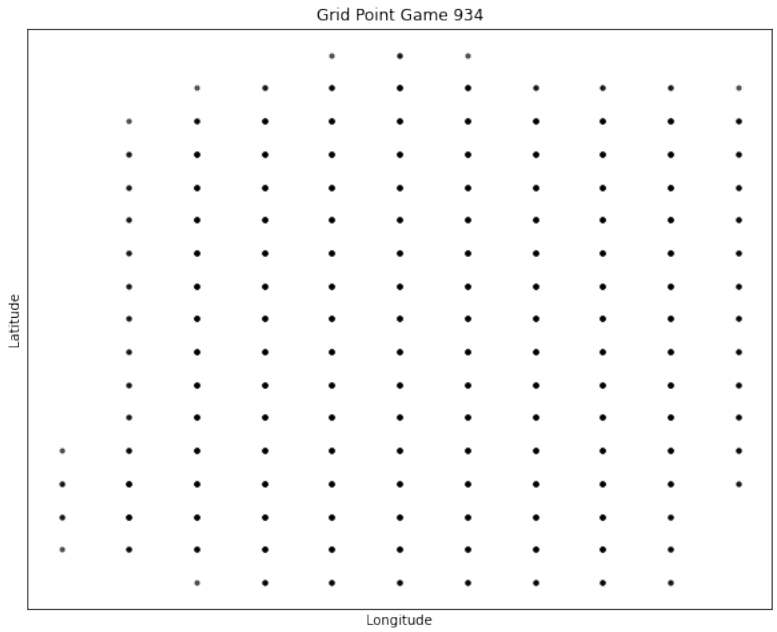


Figure 4.3: Division of playing area into cells obtained via tessellation algorithm.

area was divided into non-overlapping, gap-free cells using a tessellation algorithm [111]. Each cell corresponds to a specific area of the stadium, with a known centroid defined by latitude and longitude coordinates. Figure 4.3 shows the partition of the playing area into distinct spatial cells for Game 934 obtained with the tessellation algorithm. Only the four corner points of each cell (top-left, top-right, bottom-left, and bottom-right) are displayed in the figure to define the cell boundaries.

It should be noted that pitch dimensions were not standardized across the games analyzed, as Gaelic football pitches can vary considerably in size [26]. The tessellation algorithm was applied independently to each game using the GNSS coordinates recorded during that specific match, meaning the spatial grid was generated relative to the actual area covered in each individual game rather than a fixed reference pitch.

Player GNSS coordinates recorded during the match were mapped to the nearest cell by calculating the Euclidean distance to each cell centroid and assigning the coordinate to the closest one. Cells without assigned coordinates were excluded from the analysis, as they represent parts of the playing area or stadium not covered during play. This spatial grid enables the identification of the start and end locations

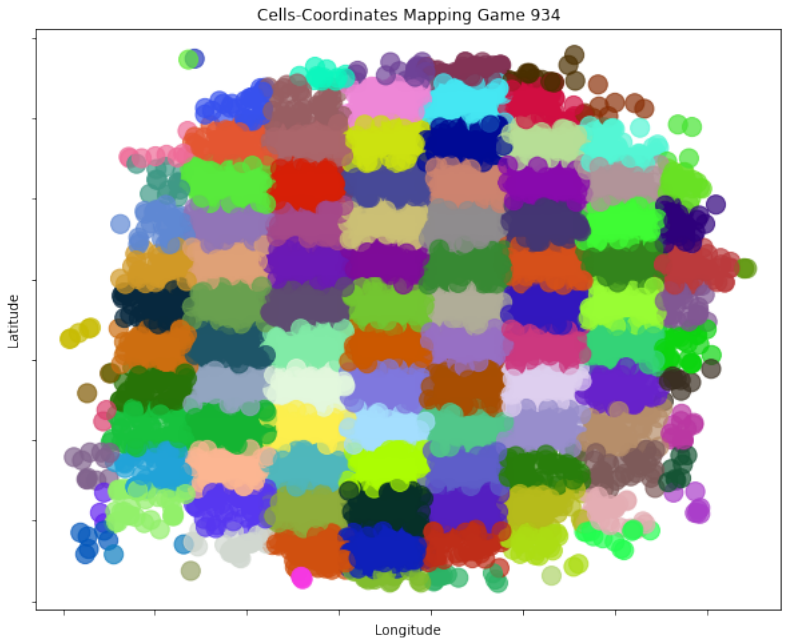


Figure 4.4: Spatial mapping coordinates-to-cells in Game 934.

for each player's action within the subdivided cells of the playing area. Grouping nearby action start or end points into the same cell permits analyzing player movements over cells rather than single data points. This approach simplifies the data without losing important information and improves the clarity and interpretability of movement patterns. Figure 4.4 illustrates the spatial mapping in our example: each coordinate is assigned to the nearest cell.

4.2.3 Describing Player Movements as a Multi-Edges Graphs

Once the playing area is divided into cells, the next step is the construction of spatio-temporal graphs. In simple terms, the playing area is divided into small zones, and players are tracked to investigate how they move between these zones during different parts of the game. For example, it can be observed whether a player frequently moves from the defensive area to midfield or often covers certain corridors on the field. This reveals where players spend most of their time and how they move around the field. By representing player movements as connections between these zones over time, coaches can better understand how their team uses space. For instance, they might discover that players gather in certain areas during specific

phases of the game or that some zones act as common routes for player transitions.

For a selected game and the relative playing area, the game is divided into fixed time intervals using rolling windows to create snapshots of spatial activity within the field's cells, represented by the Time-Window Spatial Activity Multi Graph (TWG). A rolling window is a technique for analyzing time-dependent data by calculating metrics over a fixed-size segment (window) that moves forward by a set step size. TWG is a set of spatial-activity graphs where each graph is a snapshot of activity for a timepoint w within a cell a. Each Spatial Activity Graph is a directed multi-graph $TWG_{a,w} = (N, E)$ where N is the set of nodes representing a cell of the playing area and n denotes a tuple containing the latitude and longitude of the cell's centroid $n =< lat, lng >, n \in N$. E denotes the set of edges within the graph, where each edge e indicates an action represented by a four-tuple $e =< n_{from}, n_{to}, \bar{Speed}, d, w, l, \theta, P >, e \in E$ where, n_{from} and n_{to} are the *to* and *from* nodes for the directed edge (starting and ending cells of the action), $\bar{Speed}$ is the average speed during the action, d is the duration of the action, w is the weight of the action defined as $d \cdot \bar{Speed}^{1.5}$, l the traversed distance, θ is the turning angle associated with the action and P is the player who performed the action. The multi-graph representation permits multiple edges between the same pair of spatial nodes. Each edge is described by multiple metadata attributes $(\bar{Speed}, d, w, l, \theta, P)$, which allows a richer, multi-dimensional representation of each player action.

While six attributes per edge may appear heavy, the graph remains manageable in practice because players only move between a limited number of neighboring cells, meaning most cell pairs have no connecting edge, and the overall structure remains sparse. Not all attributes are equally relevant to every coaching question: the weight w, which combines duration and speed into a measure of physical workload, is the primary attribute used in this chapter, while the others serve as optional descriptors that can be ignored depending on the task at hand.

The pseudocode for realizing the TWG is presented in Algorithm 4.1, where TW is a distinct list of time windows (for example, $[0, 5], [1, 6]$...), of 5-minute length and

1-minute step size, and A is the set of actions annotated with the cell number. The duration of the time window can be selected according to the specific objectives and temporal granularity required by the analysis

Algorithm 4.1 TWG Construction Algorithm

Require:

TW: List of time windows (for example, $[0, 5], [1, 6], \ldots$)

A: Set of player actions, each with attributes: start time, end time, start cell, end cell, player ID, average speed, duration, turning angle, distance

C: Mapping of cell IDs to centroid coordinates (latitude, longitude)

Ensure:

TWG: Set of time-windowed spatial activity graphs, each $TWG_{a,w} = (N, E)$

for all time window w in TW **do**
  for all cell a in C **do**
    Initialize empty multi-graph $TWG_{a,w} = (N, E)$
    $N \leftarrow$ set of centroids of all pitch cells (from C)
    for all action $act \in A$ **do**
      if act occurs within time window w **and** act starts or ends in cell a **then**
        $n_{from} \leftarrow C[act.start\_cell]$
        $n_{to} \leftarrow C[act.end\_cell]$
        $\bar{Speed} \leftarrow act.avg\_speed$
        $d \leftarrow act.duration$
        $w_{edge} \leftarrow d \cdot \bar{Speed}^{1.5}$
        $l \leftarrow act.distance$
        $\theta \leftarrow act.turning\_angle$
        $P \leftarrow act.player\_id$
        Add directed edge $e = (n_{from}, n_{to})$ with attributes $(\bar{Speed}, d, w_{edge}, l, \theta, P)$ to $TWG_{a,w}$
      end if
    end for
    Store $TWG_{a,w}$ in TWG
  end for
end for
return TWG

Figure 4.5 presents a simplified TWG using a small sample from Game 934. Only a few players, actions, and spatial cells (nodes 1, 2, and 3) are shown. Each colored edge represents a single action by a player moving between two cells, with action details such as player, speed, duration, distance, angle, and edge weight provided next to the edge. For example, between node 1 and node 2 there are two separate edges, each representing a distinct action. One edge corresponds to Player A, whose action lasted 3 seconds with an average speed of 5 m/s. The other edge

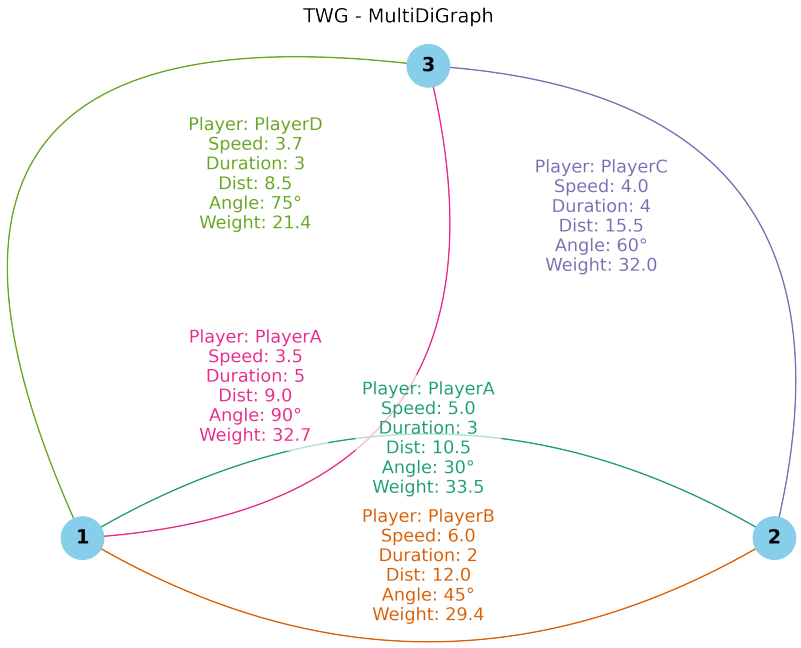


Figure 4.5: Example TWG for Game 934, using a simplified network with a few players and spatial cells.

corresponds to Player B, whose action lasted 2 seconds at an average speed of 6 m/s. This illustrates how the multi-graph representation captures multiple different movements between the same locations.

4.2.4 Aggregation into a Simpler Directed Graph

Some graph metrics, such as betweenness centrality and community detection, require a simple directed graph $TWG' = (N, E')$ without multiple parallel edges. Deriving a weighted directed graph from a multi-graph involves combining multiple edges between the same pair of nodes into a single representative edge with an aggregated weight. The steps required to transform the TWG into TWG' are the following:

1. Multiple parallel edges between nodes are aggregated by computing the average of the weight:

$$\bar{w}_{u,v} = \frac{1}{k} \sum_{i=1}^{k} w_{e_i}, \tag{4.1}$$

where the k edges e_i connect node u to node v.

2. k represent the count of edges between nodes u to node v.

3. The weight is computed to reflect both intensity and frequency, emphasizing transitions that are intense and frequent:

$$w'_{u,v} = \frac{1}{\bar{w}_{u,v} \cdot k + \epsilon}, \tag{4.2}$$

with a very small $\epsilon > 0$ for to avoid 0 division. This inverse weighting means edges with high cost ($\bar{w}_{u,v}$) and frequency k have lower weight, effectively making these edges more preferred in shortest-path calculations, node importance, and community detection analyses.

Figure 4.6 shows the aggregated weighted undirected graph (TWG') derived from the TWG for Game 934. Here, multiple player actions between the same pair of spatial cells (nodes 1, 2, and 3) have been combined into single edges. The edge labels display the computed weight values according to the aggregation and the number of original edges combined between those nodes ("k").

4.2.5 Analysis of Field Zones with Graph Metrics

From the directed graph $TWG' = (N, E')$, where N is the set of playing area's cells and $E' = \{(u, v) \in N : w'_{u,v} = \frac{1}{\bar{w}_{u,v} \cdot k + \epsilon}\}$ the set of edges, it is possible to extract graph theory metrics.

- *Node Degree (ND).* Node degree indicates the number of connections a particular playing area's cell has with other cells via player movements. A high degree means that a cell frequently serves as a starting or ending point for transitions, reflecting its involvement in multiple movement pathways. In a directed graph, there are two types of node degree:

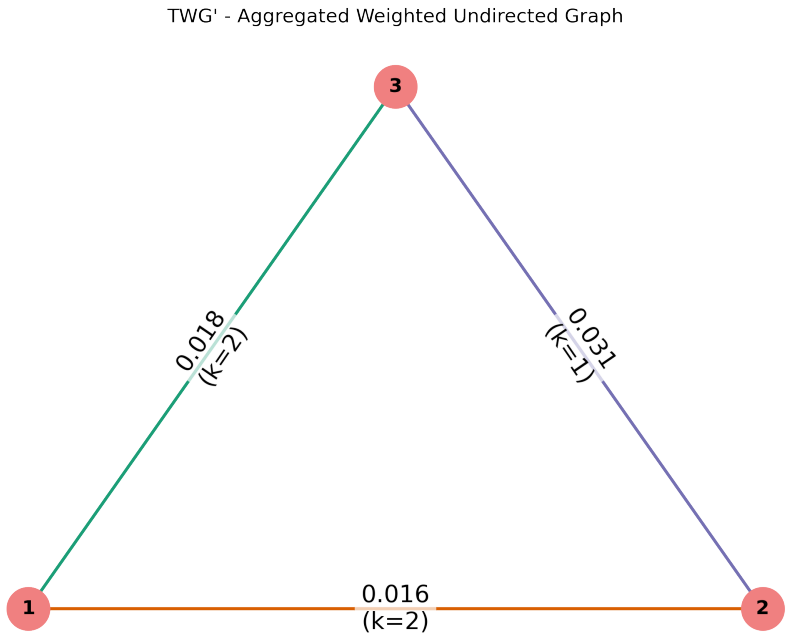


Figure 4.6: Aggregated graph (TWG') for our running example in Game 934, created by combining the edges from the TWG sample into single weighted edges between cell pairs.

– *In-degree* k_i^{in}: the number of incoming edges to cell i, representing transitions ending in cell i. This is formally defined in Equation 4.3:

$$k_i^{in} = \sum_j A_{j,i}, \tag{4.3}$$

– *Out-degree* k_i^{out}: the number of outgoing edges from cell i, representing transitions starting from cell i. Equation 4.4 formally defines k_i^{out}:

$$k_i^{out} = \sum_j A_{i,j}, \tag{4.4}$$

The *average node degree* $\bar{k}_i$ summarizes the overall connectivity level of the network, indicating the typical number of connections per cell through player movements. It is computed as the mean degree across all nodes, as shown in Equation 4.5:

$$\bar{k_i} = \frac{1}{|N|} \sum_{i=1}^{|N|} k_i^{in} + k_i^{out}, \tag{4.5}$$

where $|N|$ is the total number of cells. A higher average node degree suggests that, on average, cells are involved in more transitions, reflecting a more interconnected playing area.

- *Node Centrality.* Node centrality highlights which field cells are most influential in connecting play. Nodes with high centrality are zones that players frequently use as "bridges" to move between regions and velocity zones, making them key areas for transitions and tactical importance. In simpler terms, cells with high betweenness centrality are areas where players tend to move through frequently when transitioning between different regions. These cells are strategically important as they serve as key connectors within the team's spatial organization, reflecting regions with significant player movement activity.

  Betweenness centrality is a measure of network centrality that quantifies how often a node lies on the shortest paths between other nodes in the network [112]. The formula for calculating betweenness scores [109] is shown in Equation (4.6) where σ_{st} is the total number of shortest path going from s to t, and $\sigma_{st}(x)$ the number of those paths passing through node x.

$$BC(x) = \sum_{s \neq x \neq t} \frac{\sigma_{st}(x)}{\sigma_{st}} \tag{4.6}$$

  This centrality measure assigns to each node in the graph a score reflecting its importance (in terms of centrality and connections) in the graph.

  In Figure 4.7, it is shown the betweenness centrality computed for the first 300 seconds of Game 934. Cells with higher betweenness centrality scores (in yellow and lighter colors) indicate influential zones where many shortest paths

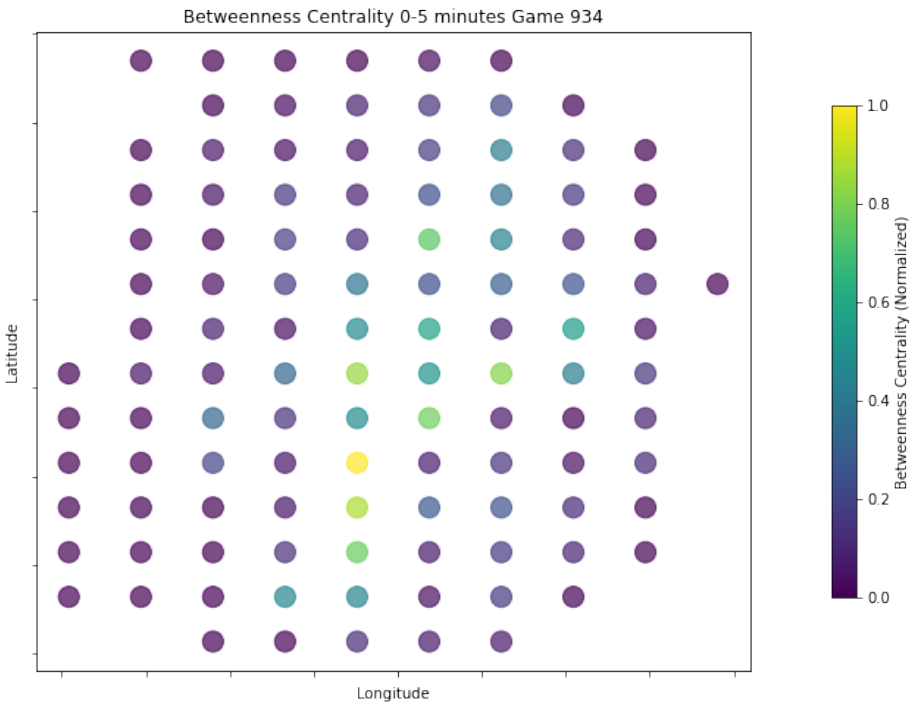


Figure 4.7: Betweenness centrality values for Game 934 during the first 5 minutes (0–300 seconds).

of player movement converge. In the first half of Game 934, the team's goal is located in the bottom-left of the image and the opposing team's goal in the top-right, meaning the direction of play runs from bottom-left to top-right. In this example, the betweenness centrality scores were highest in the defensive central area of the playing area, meaning that it was a strategic area through which players frequently passed or moved to connect different parts of the field.

- *Community Detection.* Community detection is a graph-based technique that identifies clusters of field cells frequently connected by player movements within a given time window. These communities reveal regions of the playing area formed by a group of cells, where players tend to move within and between cells more frequently than to other parts of the playing area. By applying community detection to the aggregated movement graph (TWG'), it is possible to identify spatial groupings that emerge from players' behavior during games. This offers a view of how the team is organized spatially, revealing

zones of high internal connectivity that imply more concentrated activity. For example, clusters may show areas where defenders often move between nearby zones or midfield corridors that players use to transition between different parts of the field. The number and shape of communities can change across time windows, reflecting the evolving positioning of the team. Larger communities indicate broader zones of shared activity, while smaller communities reflect more isolated areas of action.

The Louvain community detection algorithm groups data points into clusters by maximizing the density of connections within communities relative to connections between communities, as quantified by modularity using Equation (4.7). Here $A_{i,j}$ represents the weight of the edge between i and j, $k_i = \sum_j A_{i,}$ is the sum of the weights of the edges connected to node i, c_i is the community to which node i is assigned, and δ is a function such that $\delta(c_i c_j) = 1$ if node i and node j are assigned to the same community [110] and 0 otherwise.

$$M = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(c_i c_j) \tag{4.7}$$

Figure 4.8 displays the results of community detection applied to the player movement graph for Game 934 during the first 5 minutes of play. In this example, five different communities are detected, and, interestingly, no community is located in the midfield area. Instead, all groups seem to converge from the outer edges of the playing area toward the center. This is likely explained by the early phase of play, where players start from wide distributed positions and move inward as they track the ball. The midfield area in this window appears to function as a transitional corridor connecting the peripheral communities rather than a zone of sustained coordinated activity, which is consistent with the high betweenness centrality scores observed in the central area for the same game.

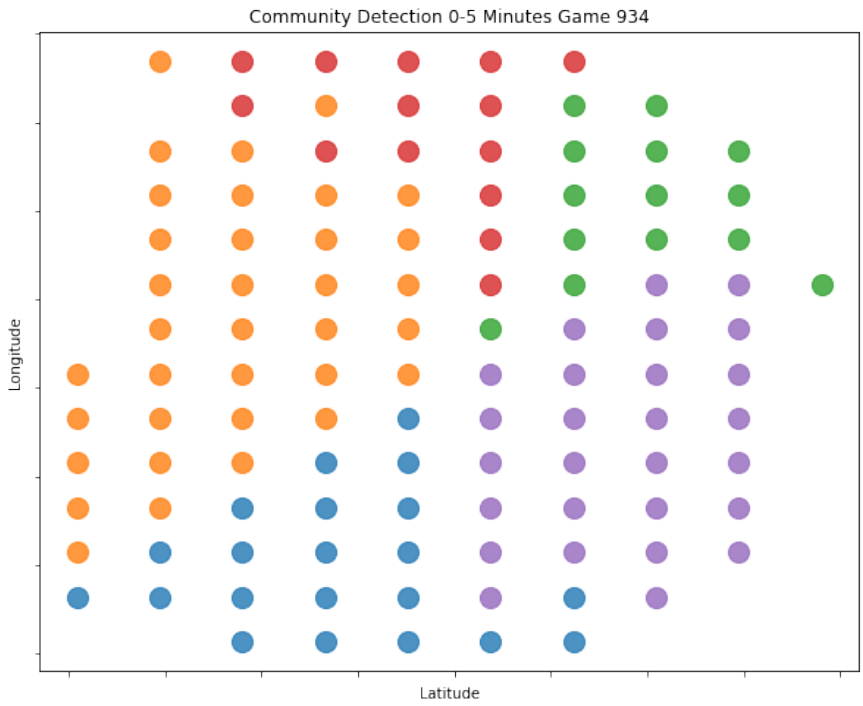


Figure 4.8: Communities detected for Game 934 during the first 5 minutes (0–300 seconds).

4.3 Results

In this section, the ability of the Time-Window Spatial Activity Graph to analyze the spatial movement of players across consecutive time windows is assessed in two different games played on diverse fields. In both cases, the analysis includes each available movement player with the exception of the goalkeeper.

Games 811 and 893 were selected as illustrative examples from the available dataset, played in two different stadiums, to demonstrate the framework across different spatial environments.

The research question asks whether graph-based representations can reveal: (1) critical transition zones, (2) spatial communities with distinct workload profiles, and (3) temporal evolution of these patterns. The two case studies address these components differently. Game 811 demonstrates all three components: betweenness centrality identifies critical transition zones, community detection reveals spatial groupings with speed profiles, and rolling 5-minute windows track temporal evolution of both transition zones and community structures. Game 893 addresses two

components: community detection with detailed workload profiling (speed, acceleration, intensity distributions) and temporal evolution of community structures between two 15-minute phases. Betweenness centrality was not computed for Game 893, thus, critical transition zones are not examined in that case study. Together, these analyses demonstrate the framework's analytical flexibility while addressing the research question's core elements.

4.3.1 Evaluation: Spatio-Temporal Analysis of Game 811

This analysis addresses all three components of the research question. Using rolling 5-minute windows across the opening 8 minutes of Game 811, it is examined: (1) how critical transition zones shift through betweenness centrality analysis, (2) how spatial communities group different playing areas and their associated speed profiles, and (3) how both transition zones and community structures evolve temporally across consecutive windows.

This analysis offers novel insights into activity in the playing area, player movement, and connectivity dynamics, contributing to a deeper understanding of the spatio-temporal complexities of Gaelic football performance. The input dataset is the Action dataset related to Game 811, comprising 13,586 actions that capture spatio-temporal coordinates and extracted features throughout the match.

Spatial Cell Mapping

After segmenting the playing area into sub-cells, every GNSS location recorded for players during the match was then associated with its nearest cell centroid. Figure 4.9 illustrates the grid points generated with the tessellation algorithm (left) and the result of assigning each coordinate to the closest cell (right).

TWG' Generation

Using the spatial cells as nodes, Temporal Window Directed Graphs (TWG') were realized to model dynamic player movements over time, generating one graph for each

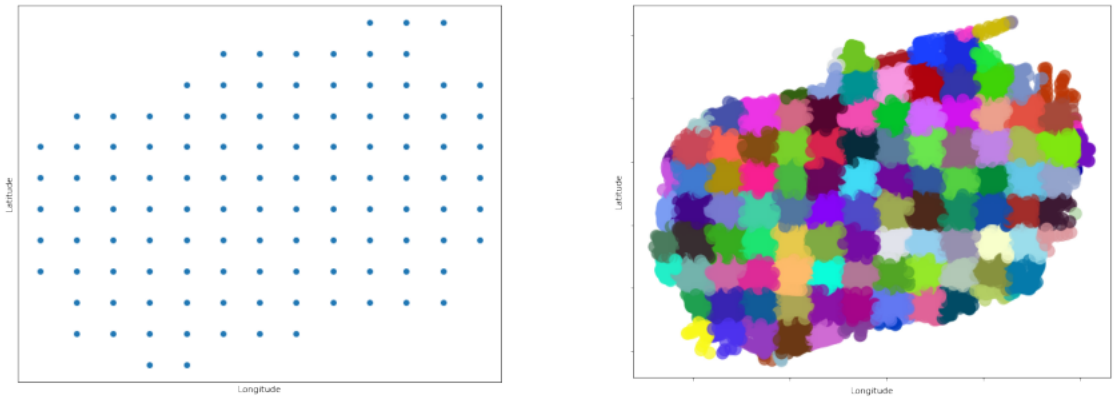


Figure 4.9: Spatial grid mapping of the pitch. **Left**: grid points obtained via tessellation algorithm. **Right**: the result of assigning each coordinate to the closest cell.

rolling time window. Following the construction of the spatio-temporal multi-graphs, each was subsequently transformed into a weighted directed graph as described in Section 5.2.1. This transformation involved merging multiple edges between spatial cell pairs into a single weighted edge that reflects both the frequency and cost (speed) of transitions. The resulting directed graphs enabled the computation of structural metrics such as betweenness centrality and community detection.

The length of the rolling window was set to 5 minutes, with a sliding size of 1 minute, resulting in the following intervals: [0,5], [1,6], etc. The 5-minute window was selected as a pragmatic compromise between temporal resolution and data volume, consistent with its established use in GPS-based match analysis in team sports, where 5-minute rolling averages have been applied to examine physical demands in professional soccer [113] [114]. Shorter windows capture finer temporal changes but contain fewer actions, which can limit the stability of graph-based metrics such as community detection, while longer windows provide more data but reduce the ability to detect tactical shifts within a half. Using a 5-minute window with a sliding size of 1 minute in the tested game, which lasted 78 minutes, results in the generation of 74 TWG'.

Identifying Key Zones Within the Playing Area Through Betweenness Centrality

The betweenness centrality of cells in the weighted directed graph (TWG'), computed for each 5-minute time window, was used to identify the most important cells where directional changes and transitions between velocity zones occur most frequently. Figure 4.10 compares betweenness centrality scores across four consecutive rolling windows: [0, 5], [1, 6], [2, 7], and [3, 8], describing the first 8 minutes of activity at the beginning of Game 811. The analysis included only cells visited by players during the time window. In this first half, the betweenness centrality scores were computed using the inverse edge weights described in Section 4.2.5, which integrate both frequency and intensity, ensuring that high-centrality cells reflect frequent paths.

The importance of locations changes during the time windows analyzed: Figure 4.10 illustrates this transition from central to peripheral areas, observable in the difference of scores between the first time window [0, 5] and two successive windows ([1, 6] and [3, 8]). In the top-right of the playing area is located the team's goal, while on the other side, bottom-left, is located the opposition team's goal.

In the first two time windows (top-left and top-right), the highest betweenness centrality scores (indicated by yellow or lighter points) are concentrated in the central corridor of the playing area. The team tends to use the middle area to connect defensive and attacking moves, with players often passing or running through these central zones. While in the third window, the central area remains important, in the fourth window (bottom-right), the highest centrality shifts toward the attacking region, particularly close to the opposition goal in the bottom-left of the playing area. This transition reflects more frequent movements through the attacking area, as the team's activity becomes focused on creating scoring opportunities. This shift is captured by the graph approach: as the number and intensity of directed movements through the attacking cells rise, these zones emerge as important connectors in the flow of play. Betweenness centrality identifies cells that act as bridges in

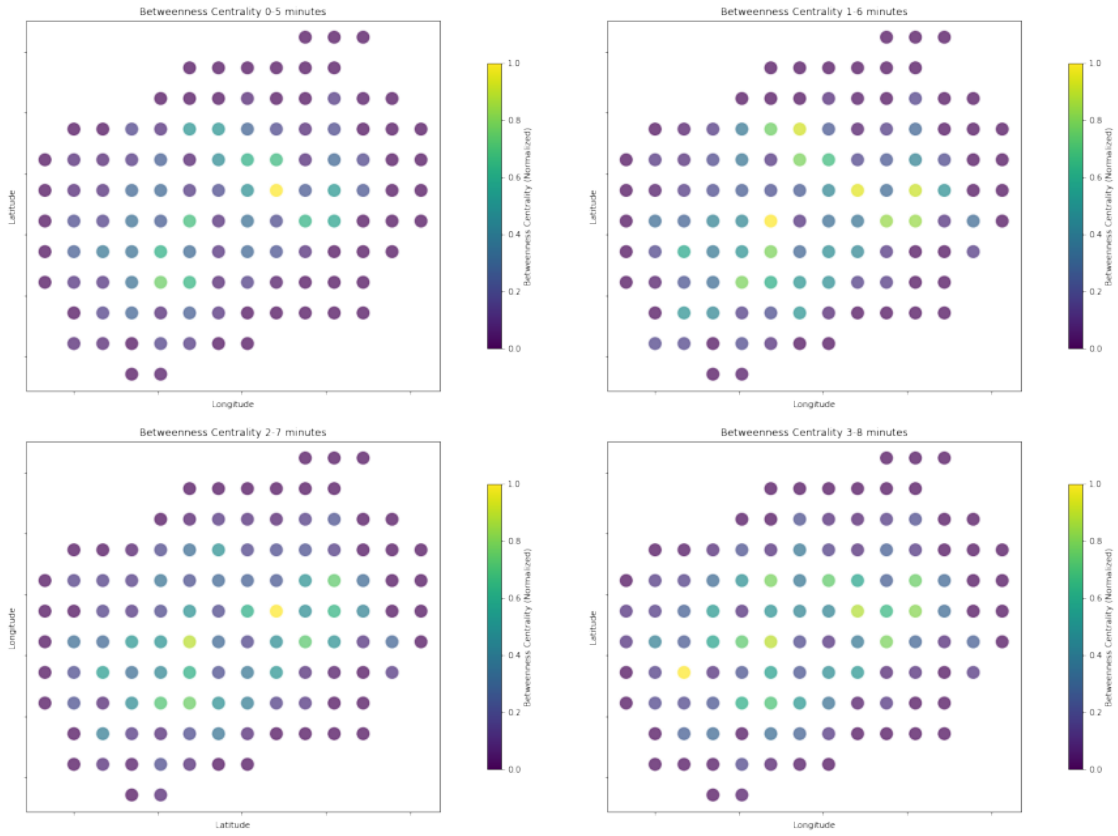


Figure 4.10: Betweenness centrality for 5-minute rolling windows: [0, 5) (**top-left**), [1, 6) (**top-right**), [2, 7) (**bottom-left**), and [3, 8) (**bottom-right**).

frequent and intense movement sequences, not only zones with high activity.

Discovering Spatial Groupings in the Playing Area

Figure 4.11 shows the results of Louvain community detection applied to the playing area's cells across four consecutive 5-minute windows in Game 811. These communities reveal clusters of adjacent pitch zones where players tend to move and change speed together.

In the first window, six distinct communities are detected. The defensive area is grouped into the same community (in brown), while the attacking area is fragmented into three different communities (yellow, green, and red). The red community (external attacking side) records the highest first quartile speed (2.4 m/s), average (3.6 m/s), and third quartile (4.9 m/s). For a coach, this means the red zone during this window was the scene of high-speed movements. In contrast, yellow and brown communities cover areas with the slowest speeds.

Four communities emerge in the [1,6) window: approximately divisible in defensive-left (red), defensive-right (blue), offensive-left (green), and offensive-right (yellow). Defensive players move as a single cohesive block and a small group on the right. Similar to the previous window, the offensive-left community shows the highest speed statistics, operating as a corridor for fast transitions.

In the third window, on the bottom-left of Figure 4.11, five communities are now identified by the algorithm. The defensive area is now divided into three groups: center-right (blue), center (purple), and center-left (red), reflecting less grouped movements of the players.

Five communities persisted in the last window, but the yellow community now represents the main region of high-speed movements, with higher speed quartiles and averages. Its position signals a phase of increased attacking movements, confirmed by the fact that the slowest activities are concentrated in the red community (part of the defensive area). These changes in groups, movement, and speed let coaches understand when and where the team is pushing forward to attack or staying more compact to protect their own goal.

Discussion

This case study addressed all three research question components through rolling 5-minute window analysis.

Critical transition zones: Betweenness centrality quantified shifting zone importance from the central corridor (windows [0,5) and [1,6)) to the attacking region (window [3,8)), demonstrating tactical evolution from build-up to offensive pressure.

Spatial communities with workload profiles: Community detection revealed distinct speed characteristics across zones, with the external attacking side (red, 3.6 m/s average) exhibiting highest speeds in window [0,5), the offensive-left corridor (green) in window [1,6), and the center-right offensive area (yellow) in window [3,8).

Temporal evolution: Both transition zones and community structures evolved

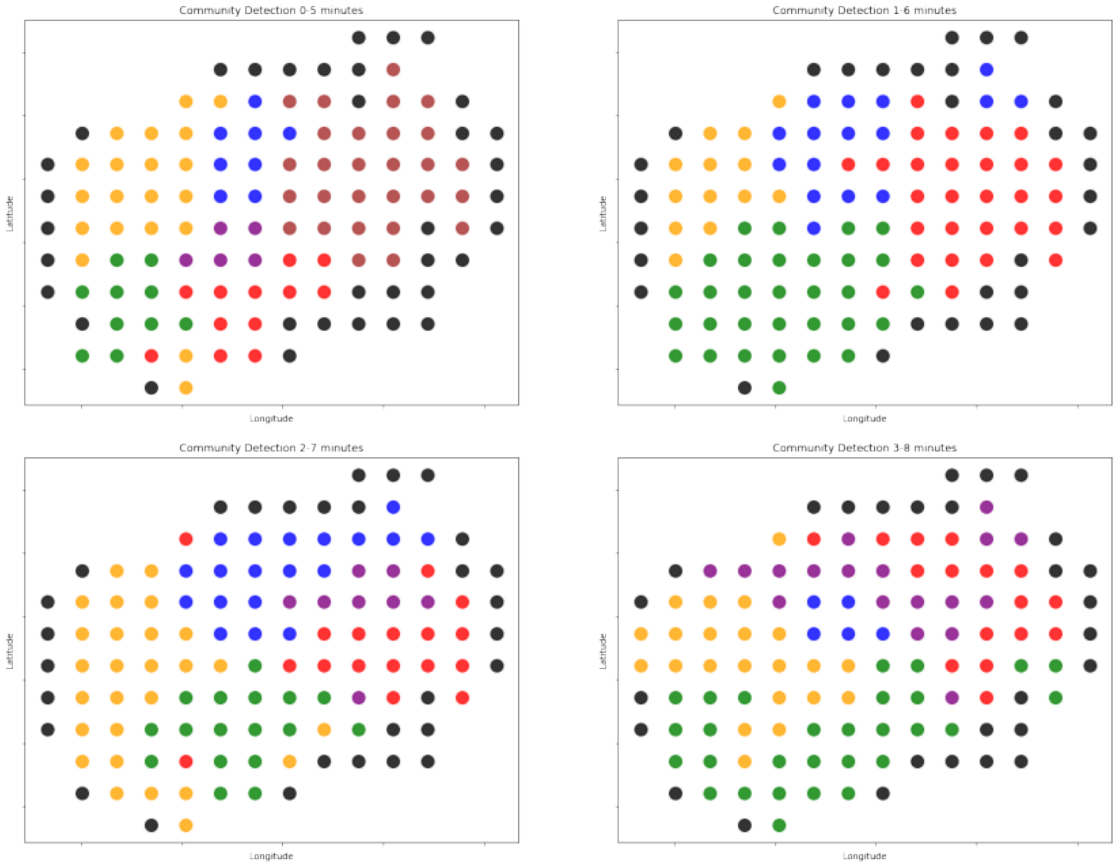


Figure 4.11: Community detection for 5-minute rolling windows: [0, 5) (**top-left**), [1, 6) (**top-right**), [2, 7) (**bottom-left**), and [3, 8) (**bottom-right**). Black points represent coordinates not covered by players in the analyzed time window.

dynamically: betweenness centrality shifted from central to attacking regions, community organization reorganized from unified defensive (brown) to fragmented defensive zones, and high-speed corridors relocated sequentially across different attacking areas (red → green → yellow).

4.3.2 Evaluation: Spatio-Temporal Analysis of Game 893

This case study examines the first 30 minutes of Game 893, divided into two distinct and consecutive non-overlapping time windows: 0–15 minutes and 15–30 minutes. The previous analysis used rolling windows, whereas this approach compares fixed intervals to assess changes between separate phases of play.

Fixed 15-minute intervals are a well-established unit in team sport match analysis for examining temporal changes in physical and technical performance across distinct phases of play [115], [116]. The choice here was deliberate: rather than capturing short-term fluctuations, the aim was to compare broad spatial community structures between early and mid-match phases. Aggregating actions over each window is therefore an intentional design decision, with the acknowledged trade-off that within-phase variations are not captured at this resolution.

This analysis addresses two components of the research question: spatial communities with distinct workload profiles and their temporal evolution. Using two consecutive 15-minute windows (0-15 min and 15-30 min), it is examined: (1) how spatial communities group different playing areas and their associated physical characteristics (speed, acceleration, intensity distributions), and (2) how community structures and workload profiles evolve between early and mid-match phases. Betweenness centrality is not computed in this case study, as the focus here is on community-level workload profiling rather than transition zone analysis, providing a complementary perspective to the Game 811 case study.

The match (Game 893) selected for this case study was a county-level male Gaelic football match with 16 players, which lasted for 75 minutes. The action dataset of this match consists of 12688 records.

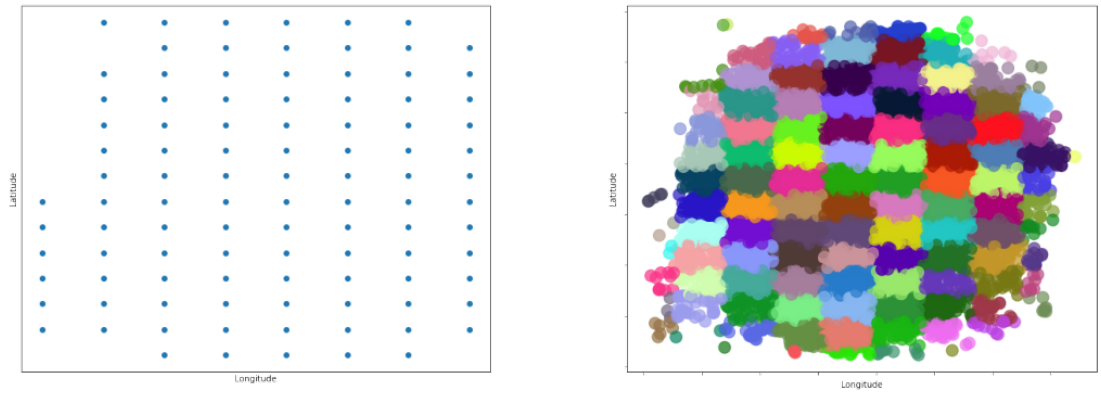


Figure 4.12: Spatial grid mapping of the pitch. **Left**: grid points obtained via tessellation algorithm. **Right**: the result of the association of players' coordinates to their closest grid point.

Spatial Cell Mapping

The initial step involves creating a grid of the playing area through a tessellation algorithm, which divides the space into non-overlapping, gap-free cells. Player coordinates are then assigned to the nearest cell, as illustrated in Figure 4.12.

TWG' Generation

In this case study, analysis focused on players' actions during two consecutive non-overlapping time windows: the first 15 minutes of the game and from the 15th to the 30th minute. Using the action dataset, we then created the multi-graphs TWGs (one for each time window) and the subsequent directed simple graphs TWG's, aggregating multiple edges between nodes following the logic explained in Section 5.2.1. The edges between the nodes take into account the average speed of the actions starting in a node and ending in the other, in addition to the duration and the frequency (Section 5.2.1).

Spatial Analysis of Actions in the 0–15 Minute Time Window

Spatial communities detected through Louvain community detection within the first 15 minutes of play were analyzed by visualizing the spatial distribution of players' movements in Figure 4.13 and summarizing their running activity within the de-

tected communities in Table 4.1. The visualization in Figure 4.13 displays TWG' with nodes colored according to communities detected via the Louvain algorithm, edges colored and weighted to reflect their average speed, and node sizes scaled by their degree. Table 4.1 details the number of playing area cells and unique players per community, along with average and interquartile range for speed and acceleration, high-acceleration events, and the proportions of distinct action intensities categorized as standing/walking, jogging, running, and high-intensity running/sprinting.

Among the detected communities, Community 1 stands out as the largest with 37 cells and all 14 players traversing these cells, followed by Community 2 (28 cells, 14 players), Community 7 (18 cells, 14 players), and Community 0 (10 cells, 13 players). These four communities collectively represent the entire playing area. Considering the goalkeeper's position, we can identify Community 2 as corresponding to the central and defensive center-right regions of the playing area, Community 7 as covering the defensive-left and left regions, Community 0 as representing the offensive left zone, and Community 1 as encompassing the offensive and central offensive right regions. Smaller communities have many fewer cells and involve fewer unique players, indicating more isolated zones of action. Figure 4.13 illustrates that the majority of actions occur in the central region of the playing area, with slightly more activity on the left side compared to the right, as suggested by the density of edges and the sizes of the nodes. The offensive regions also show increased actions at higher speeds, as indicated by the lighter color of the edges. In addition, the directions of these actions seem to follow diagonal trajectories moving from the defensive toward the offensive zones, suggesting that most movements at higher speeds involve forward, diagonal runs during attacking phases. Community 1 (offensive and central offensive right) is characterized by the largest spatial extent and player involvement, the highest count of high-acceleration events, and a slightly higher proportion of running actions. Community 2 (central and defensive center-right) balances moderate average speeds with substantial acceleration variability and high acceleration events, reflecting the presence of mixed-intensity actions. Community 7 (defensive-left and left

Table 4.1: Summary of spatial communities with key physical features for the time window 0-15 minutes. For each community, the number of nodes and unique players, average and spread of speed (AvgSpd, SpdIQR) and acceleration (AvgAcc, AccIQR), total high-acceleration events (HighAcc), and proportions of key action intensities: combined Standing/Walking (Stand/Walk), Jogging (Jog), Running (Run), and High-Intensity Running plus Sprint (HIRun_Sprint) are shown. Values represent means or proportions per community.

Comm.	Nodes	Players	AvgSpd (m/s)	SpdIQR (m/s)	AvgAcc (m/s^2)	AccIQR (m/s^2)	HighAcc (count)	Stand/Walk %	Jog %	Run %	HIRun_Sprint %
0	10	13	2.88	2.30	1.71	0.77	25	0.34	0.41	0.16	0.07
1	37	14	2.93	2.50	1.81	0.80	82	0.32	0.42	0.19	0.06
2	28	14	2.74	2.40	1.70	0.90	74	0.36	0.39	0.18	0.05
3	2	3	1.94	1.50	1.75	0.99	1	0.60	0.40	0.00	0.00
4	1	1	5.50	0.90	2.18	0.03	1	0.00	0.00	0.50	0.50
5	2	3	1.81	1.10	1.41	1.03	1	0.66	0.33	0.00	0.00
6	1	1	0.10	0.00	0.40	0.00	0	1.00	0.00	0.00	0.00
7	18	14	3.02	2.60	1.80	0.82	43	0.30	0.41	0.19	0.08
8	1	1	1.50	0.00	1.49	0.00	0	1.00	0.00	0.00	0.00

areas) exhibits the highest average speed and proportion of high-speed running or sprinting actions among the four communities, indicating frequent high-movement actions typical of defensive or counter-attacking phases. Community 0 (offensive left) exhibits similar average speed and acceleration values to Communities 1 and 2, with a proportion (7%) of high-speed running and sprinting actions. Furthermore, it demonstrates a rate of high-acceleration events at approximately 0.40 per cell, indicating that despite its smaller spatial size, this community still represents cells characterized by intense player activity.

It is acknowledged that 15-minute communities are limited by design and reveal overall spatial organization rather than fine-grained tactical shifts. The more informative signal at this resolution comes from the edge colors in Figure 4.13: the concentration of lighter edges along the left corridor and central offensive corridor indicates that high-speed activity is not uniformly distributed but systematically channelled through specific areas of the playing area. This suggests the team's attacking play during this phase is built predominantly around left-side and central forward runs, while defensive zones show consistently lower speeds, reflecting a more controlled movement profile in those areas.

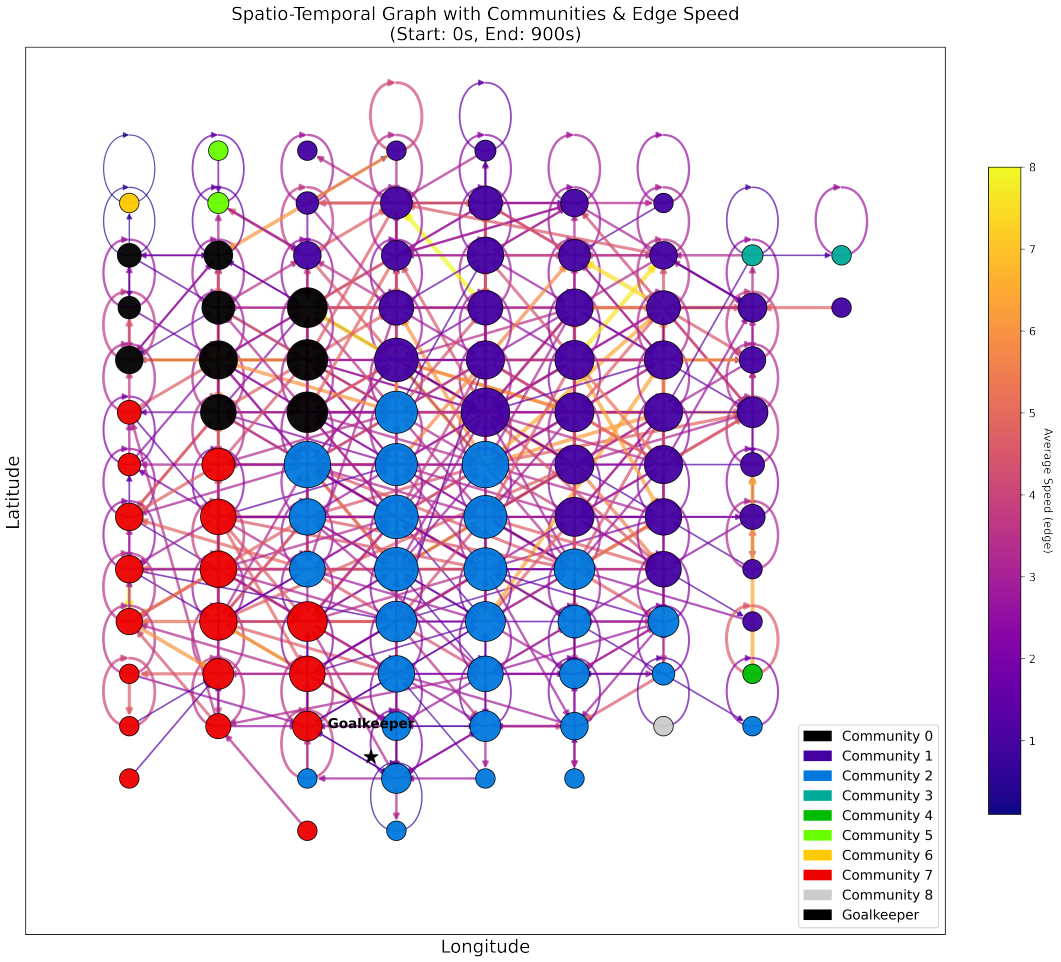


Figure 4.13: Spatio-temporal graph depicting detected communities during the first 15 minutes of the game (0, 900 seconds). Node colors represent communities, edge colors indicate average speed, and node sizes correspond to their degree. The label 'goalkeeper' represents the average position of the goalkeeper in the time window.

Spatial Analysis of Actions in the 15–30 Minute Time Window

This section uses the TWG' to analyze the spatial running performance of players during the 15th to 30th minute time window of the selected game. Figure 4.14 shows the TWG' for the analyzed window. Since only the cells traversed during this time window are inserted as nodes, the graph looks slightly different from the previous graph or the spatial grid mapping of the field (where the entire set of latitude and longitude coordinates covered during the entire match was used).

Louvain community detection identified a lower number of communities (6) compared to the previous time window, distinguishing 2 main communities: Community 0 (midfield and offensive regions of the playing area) and Community 1 (defensive and left regions). The majority of traversed cells are situated in the midfield and center-offensive regions. There is a notable presence of lighter colored edges on the left side of the pitch, in both defensive and offensive zones, indicating a higher frequency of actions at higher speeds in these zones.

Community 0 is the largest and most active spatial cluster, with 39 cells and involving all 14 players who traversed it, indicating its central role in overall player movement. This community exhibits a relatively high average speed of 3.0 m/s, coupled with a broad speed interquartile range of 2.40 m/s, which reflects a wide variety of movement intensities consistent with dynamic and physically demanding regions, such as midfield or offensive zones. In terms of acceleration, Community 0 leads present a total of 115 high-acceleration events, with an average of 2.94 acceleration events per cell. This highlights frequent bursts of speed and intense activity throughout this zone. Looking at the distribution of action intensities, there is a balanced mix: about 30.5 of the actions are combined standing/walking, 42.4% are jogging, 20.3% are running, and 6.7% are high-intensity running or sprinting. While players spend significant time in moderate movement phases (jogging), they also engage regularly in high-speed and intense effort actions within this community.

Community 1 is closely comparable in size to Community 0, with 37 nodes and the same number of participating players (14). Speed metrics for Community 1 are

Table 4.2: Summary of spatial communities with key physical features for the time window 15-30 minutes. For each community, the number of nodes and unique players, average and spread of speed (AvgSpd, SpdIQR) and acceleration (AvgAcc, AccIQR), total high-acceleration events (HighAcc), and proportions of key action intensities: combined Standing/Walking (Stand/Walk), Jogging (Jog), Running (Run), and High-Intensity Running plus Sprint (HIRun_Sprint) are shown. Values represent means or proportions per community.

Comm.	Nodes	Players	AvgSpd (m/s)	SpdIQR (m/s)	AvgAcc (m/s^2)	AccIQR (m/s^2)	HighAcc (count)	Stand/Walk %	Jog %	Run %	HIRun_Sprint %
0	39	14	3.00	2.40	1.80	0.78	115	0.30	0.42	0.20	0.07
1	37	14	2.75	2.10	1.70	0.89	82	0.37	0.39	0.16	0.06
2	4	7	1.93	1.70	1.45	1.03	5	0.59	0.35	0.05	0.00
3	1	3	1.72	0.52	1.68	1.04	1	0.75	0.25	0.00	0.00
4	1	1	4.1	0	1.2	0	0	0.00	0.00	1.00	0.00
5	9	14	2.62	2.32	1.65	0.97	22	0.41	0.37	0.16	0.05

slightly lower, with an average speed of 2.75 m/s and an interquartile range of 2.10 m/s, reflecting less variability in movement intensity compared to Community 0. Even the number of acceleration events is lower (82), with an average of 2.22 per cell.

Action intensity proportions for this community, representing the defensive and left side of the playing area, show a higher frequency of lower-intensity movements, with approximately 37.5% of actions in standing/walking. Jogging accounts for 39.8%, running 16.6%, and high-intensity running or sprinting 6.0%.

In this time window, midfield and offensive regions exhibit higher average speeds, greater speed variability, and more frequent high-acceleration events, indicating dynamic and intense activity. The defensive and left zones show slightly lower speeds and fewer intense actions at higher speeds, with the vast majority happening on the left side.

Discussion

This case study addressed two research question components: spatial communities with distinct workload profiles and their temporal evolution. The spatial running activity across the playing area in the two time windows (0–15 minutes and 15–30 minutes) reveals both consistent patterns and changes that can inform coaching perspectives.

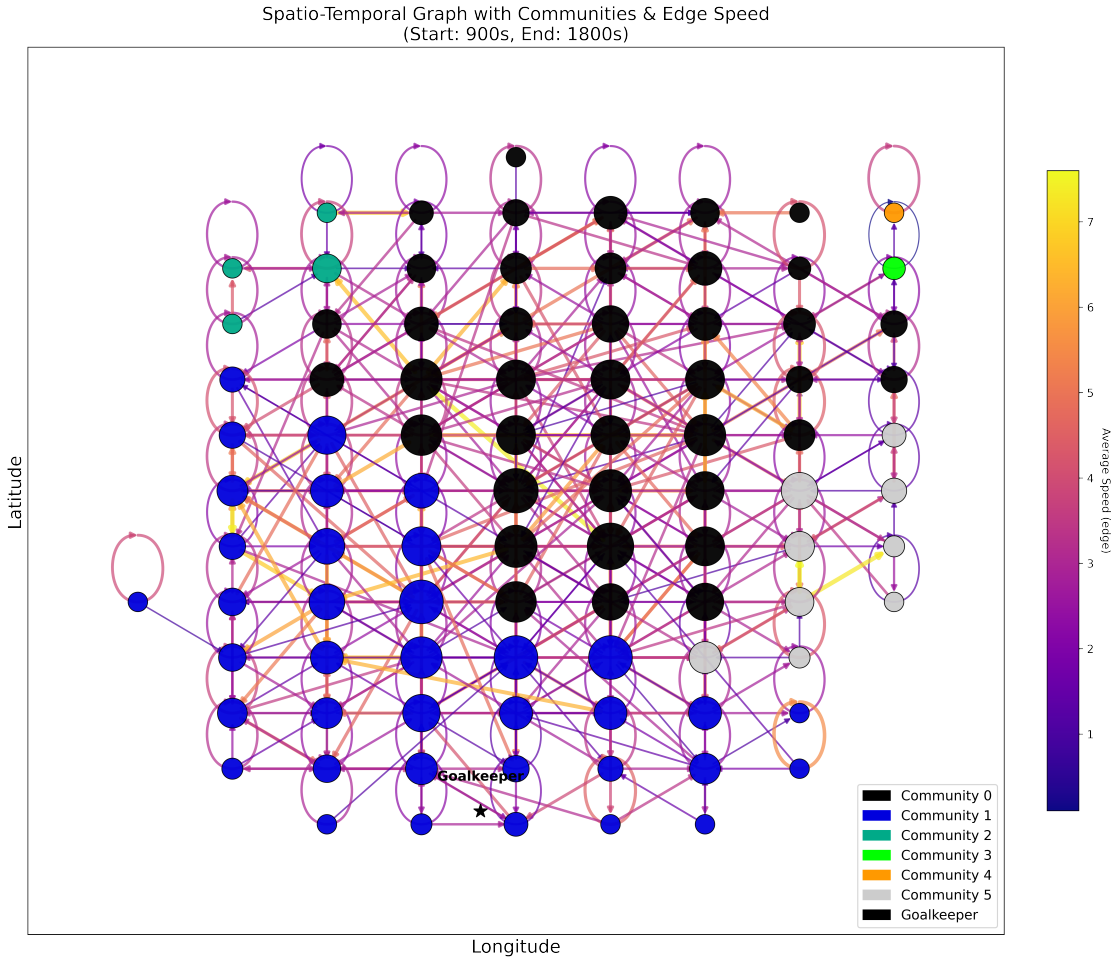


Figure 4.14: Spatio-temporal graph depicting detected communities during the time window [900, 1800 seconds), corresponding to the interval from the 15th to the 30th minute of the game. Node colors represent communities, edge colors indicate average speed, and node sizes correspond to their degree.

In the midfield and offensive regions (Communities 1 and 0 in Figure 4.13 and Community 0 in Figure 4.14), both time windows show similarly high average speeds around 2.9 to 3.0 m/s, with wide speed variability ($\tilde{2}$.4 m/s) and frequent high-acceleration events. This pattern highlights these central and attacking corridors as regions where players alternate between moderate-paced jogging and explosive sprints or accelerations.

The defensive zones (Community 2 in Figure 4.13 and Community 1 in Figure 4.14) observed lower average speed (around 2.7 to 2.8 m/s) and fewer acceleration bursts, indicating a generally more measured pace.

On the left side of the playing area (Community 7 in Figure 4.13 and the left portion of Community 1 in Figure 4.14), the presence of high-speed actions persists across both windows. This suggests that the left side remains an active zone for quick movements, possibly related to counterattacks, overlapping runs, or defensive pressing actions.

When considering the right and central defensive zones, the data show less pronounced high-speed activity compared to other sides and slightly more moderate intensity overall. The right side's activity tends to cluster around running and jogging, with fewer explosive acceleration events, especially in the first 15 minutes. However, during the early phase, the offensive-right corridor (Community 1 in Figure 4.13) displayed one of the widest spatial coverages and the highest number of high-acceleration events. As the match progressed, activity in this corridor became less intense, both in terms of frequency and speed.

Between the two windows, the main evolution is a simplification in spatial community structure: from multiple smaller communities in the first window to fewer, larger communities in the second. This may indicate an extension of player movements into broader zones, suggesting reduced distance between lines of players and more uniform movements. Despite this spatial simplification, the physical activity metrics per pitch area remain largely consistent between the two time windows.

For coaches, these findings reinforce that the midfield and offensive regions de-

mand sustained high intensity and variability in movement, requiring endurance for repeated high-speed efforts. The left side's persistent bursts of high-speed actions suggest high loads, necessitating specialized conditioning and tactical drills for players operating there. Defensive zones overall show a more tempered pace but intermittently require sudden high-intensity actions, highlighting the importance of agility and rapid reactive capacity in defensive roles. Last, the analysis indicates that the right side of the playing area is less frequently occupied by players, suggesting potential opportunities for more effective utilization.

4.4 Summary

This chapter introduced and validated the Time-Window Spatial Activity Graph framework as a novel graph-based method to analyze and interpret the spatio-temporal dynamics of Gaelic football players during competitive matches. By transforming detailed player action datasets into spatial graphs, the TWG approach enables a better understanding of player movements, speed variations, and spatial connectivity across the playing field over time.

The methodological contribution lies in segmenting the playing area into spatial cells and constructing directed multi-graphs within time windows, embedding metadata on the edges. This representation supports graph-theoretic analysis, including betweenness centrality and Louvain community detection, which reveal strategic zones and clusters of coordinated activity.

In two case studies involving male county-level Gaelic football matches, the framework demonstrated its capability to analyze player dynamics and spatial strategies. Betweenness centrality revealed the shifting importance of cells over time, highlighting zones frequently acting as connectors for player transitions. Community detection identified groups of spatial cells where players tend to move together more often than in other areas on the field and described connectivity patterns with running workload metrics. In the two case studies presented, the midfield and offensive zones exhibited elevated speeds and frequent high-intensity efforts, the defensive

area showed reduced speeds, and the left side was characterized by high-speed actions. These observations should be interpreted as illustrative of the framework's analytical capability rather than as established positional trends, given the limited number of games analyzed.

In summary, the TWG framework advances the quantitative analysis of team sport performance by integrating spatial, temporal, and physical dimensions into a comprehensive graph-based model.

The present chapter should be understood as a proof-of-concept demonstration of the TWG framework rather than a basis for establishing generalized spatial patterns in Gaelic football. The two case studies involve a single team across a small number of inter-county matches, and the observed spatial patterns may reflect team-specific tactical organization, opposition-dependent behavior, or match-specific contextual factors rather than universal characteristics of the sport. Comparisons with similar graph-based spatial frameworks in other sports are limited by the novelty of this approach: while spatio-temporal graph models have been applied to movement analysis in domains such as urban mobility [79] and environmental activity [77], direct equivalents applied to invasion field team sport performance analysis have not been established in the literature, reinforcing the methodological contribution of this work. Several limitations should be acknowledged. Pitch dimensions varied across games, and the tessellation grid was derived independently per match, meaning spatial cells are not directly comparable across fixtures. The analysis excluded the goalkeeper, and the rolling window parameters were selected pragmatically rather than validated against alternative configurations. Future work should apply the framework across a larger sample of games, teams, and competitive levels to assess whether the spatial patterns identified here are consistent or context-dependent. Systematic comparison of cell size parameters and window lengths would also strengthen the methodological foundation, and extension to multi-player and opposition-aware graph representations would substantially increase tactical interpretability.

Chapter 5

Forecasting Transitions in Metabolic Workload

The work presented in this chapter has been accepted for publication in:

- Antonini Valerio, Michael Scriney, and Mark Roantree. "Forecasting High-Intensity Activity in Elite Gaelic Football: A Machine Learning Approach Using Wearable Sensor Data and Metabolic Work Ratio." Accepted at the *3rd International Sports Analytics Conference and Exhibition (ISACE)*, June 2026. To appear in conference proceedings.

5.1 Introduction

While Chapter 4 provided a retrospective analysis of player movements using spatial-temporal graphs, a critical gap remains in managing athlete performance during matches. Current approaches mostly provide retrospective descriptions of player activity [23], [24], [54], without the ability to anticipate changes in workload that could affect players' physical performance and fatigue levels. This lack of predictive methods may limit coaches and practitioners from optimizing players' time management during games. This chapter addresses this gap by introducing a predictive

machine learning framework designed to forecast short-term transitions in metabolic power output. By leveraging the Metabolic Work Ratio (MWR), a metabolic power-based metric introduced in this thesis, which captures the balance between exertion and recovery within fixed time windows, this chapter moves beyond descriptive analytics toward predictions using machine learning models. This work presents a proof-of-concept framework demonstrating the possibility and accuracy of predicting metabolic load transitions, offering a foundation for technology systems to implement this functionality.

5.1.1 Metabolic Power Framework

The metabolic power model is a theoretical approach designed to estimate the metabolic energy demands of accelerative and decelerative movements in team sports [80], [117]. It operates on the principle that running with acceleration (deceleration) on flat terrain is energetically comparable to running uphill (downhill) at a constant speed, with the slope determined by the forward acceleration [80]. Given the well-established energy cost of uphill running, it is possible to translate the forward acceleration during sprint running into the equivalent up-slope and obtain the corresponding energy cost [80]. This energy cost is combined with the instantaneous forward speed to compute the metabolic power [80]. This theoretical model has been applied to soccer to evaluate match performance in top-level players through video analysis and GNSS data, enabling the assessment of instantaneous metabolic power and redefining the concept of high-intensity effort based on metabolic power rather than speed alone [81], [118]. Metabolic power is defined as the required rate of ATP hydrolysis to perform the muscular work during running, representing the net metabolic demands incurred by locomotion [119]. It is calculated as the product of instantaneous energy cost (expressed in $J \cdot kg^{-1} \cdot m^{-1}$) and speed ($m \cdot s^{-1}$), and is typically expressed in $W \cdot kg^{-1}$, indicating the amount of energy required per second per kilogram [120]. The formula that calculates the metabolic power MP was taken from [80] and is derived from Equations 5.4 to 5.2. The computation proceeds in

three steps: first, the forward acceleration is used to compute the equivalent slope ES and equivalent mass EM; these are then combined to obtain the energy cost EC; finally, EC is multiplied by speed to yield metabolic power MP.

$$ES = \tan\left(90 - \arctan\left(\frac{g}{a_f}\right)\right) \tag{5.1}$$

ES is the equivalent slope [80], representing the incline that would produce the same energetic cost as the observed acceleration. Here $g \approx 9.81\ m \cdot s^{-2}$ is the gravitational acceleration and a_f is the forward acceleration ($m \cdot s^{-2}$).

$$EM = \sqrt{1 + \left(\frac{a_f}{g}\right)^2} \tag{5.2}$$

EM is the equivalent mass [81], reflecting the additional inertial load imposed on the athlete by acceleration. It scales the energy cost to account for the effort of changing speed.

$$EC = \left(155.4\,ES^5 - 30.4\,ES^4 - 43.3\,ES^3 + 46.3\,ES^2 + 19.5\,ES + 3.6\right) \cdot EM \tag{5.3}$$

EC is the energy cost of accelerated running ($J \cdot kg^{-1} \cdot m^{-1}$) [80]. The polynomial coefficients were empirically derived by di Prampero et al. [80] from the known energy cost of uphill running at constant speed, mapping the equivalent slope to a metabolic cost per unit distance. EM scales this cost to account for the inertial load of acceleration.

$$MP = EC \cdot v \tag{5.4}$$

MP is the metabolic power ($W \cdot kg^{-1}$), computed as the product of the energy

cost EC and the instantaneous running speed v $(m \cdot s^{-1})$ [80].

Gaelic football is considered a high-demand sport, a high-speed game with few moments of pause and a lot of consecutive stops and sprints performed by players. Elite players typically cover between 8,160 ± 1,482 m and 9,222 ± 1,588 m during match play. Players perform 166–184 accelerations and cover 445–524 m at very high speeds (> 22 km/h), with high-speed running (≥ 17 km/h) accounting for 1,677 ± 419 m to 1,731 ± 659 m, depending on positional demands [121]. This nature of the game makes it one of the most strenuous sports. The stop-start nature of soccer match-play suggests that focusing solely on velocity thresholds may provide an incomplete analysis. Therefore, it is necessary to account for accelerations, decelerations, and the energy expenditure associated with changes in velocity [84].

Significant associations between metabolic power variables and objective measures of aerobic fitness were reported in elite Gaelic football players in [47]. The study highlights significant associations between metabolic power variables (among which is high metabolic load distance) and objective measures of aerobic fitness, such as VO2max and velocity at VO2max (the maximum speed athletes can sustain while using their maximal oxygen uptake) [47]. These associations suggest that average metabolic power and high metabolic load distance offer a practical indicator for monitoring aerobic fitness responses to training and match play in elite Gaelic football [47].

Metabolic power variables have been explored in elite Gaelic football, with research demonstrating their utility in understanding the positional and temporal demands of match play. For instance, midfielders were shown to have the highest activity profiles, covering greater total and high-speed distances and expending more energy compared to other positions. Additionally, significant reductions in metabolic power and running-based variables were observed as matches progressed, emphasizing the dynamic nature of match demands [48].

Nevertheless, the metabolic power model carries important limitations that are relevant when applied in the Gaelic football context. The model was originally

developed and validated using soccer data [81], and its accuracy in estimating energy expenditure assumes that locomotor cost is the dominant contributor to total energetic demand. In Gaelic football, this assumption may be weakened by the substantial energetic contribution of contact-based actions (including tackling and blocking), which involve significant muscular effort but produce little or no displacement detectable by GNSS sensors. These demands are not captured by metabolic power calculations, meaning that the true physiological load may be underestimated, particularly for positions with high contact involvement. These limitations should be considered when interpreting metabolic power-based metrics in the Gaelic football context.

This chapter addresses the gap between descriptive workload monitoring and the predictive capability needed to support real-time coaching decisions. The specific aims are:

1. The introduction of the Metabolic Work Ratio (MWR), a player-specific metric quantifying the balance between high-intensity metabolic demand and recovery within fixed time windows.

2. The development of a machine learning framework that forecasts short-term transitions in metabolic workload using features derived from GNSS and IMU sensor data.

3. The assessment of whether accurate prediction of imminent workload changes is achievable from data collected during elite Gaelic football matches, as a proof-of-concept for future real-time implementation.

5.2 Methodology

This section describes the methodology used to predict player metabolic workload using data from wearable sensors. Data were collected using the STATSports Apex 10 Hz GNSS and 100Hz IMU device, as described in Section 3.2.1. Our approach

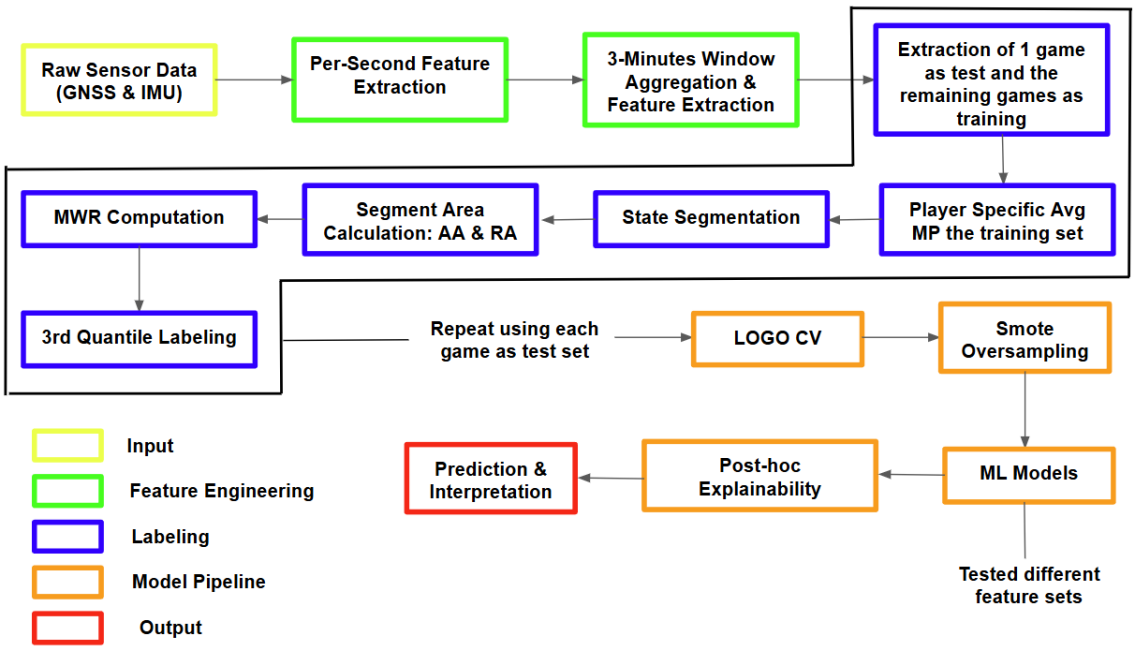


Figure 5.1: Methodology diagram illustrating the sequential data processing and analysis steps.

combines GNSS and IMU data with the Metabolic Work Ratio (MWR), a workload metric introduced in this thesis, which captures the balance of energetic effort and recovery within fixed time windows. To the best of the candidate's knowledge, no equivalent metric combining player-specific metabolic baselines with active and recovery area ratios in this form has been reported in the existing literature. The methodology consists of two main parts: first, feature engineering and metric construction, which transform raw sensor data into workload indicators, and second, a machine learning pipeline that predicts workload changes using these engineered features.

This section first addresses data preprocessing and feature extraction from raw sensor data, followed by the MWR metric computation and workload labeling methodology. The modeling pipeline is subsequently introduced, covering cross-validation, class imbalance management, and model interpretation. The approach is summarized in Figure 5.1.

For this study, features were extracted from 3 minutes of GNSS and IMU data during games to describe player activity within that time window. These features were then combined with cumulative features representing the player's activity from

the start of the game. Using this feature set, the prediction targets whether the MWR in the next time window will exceed or fall below the player's average MWR from previous games.

5.2.1 Input Data

This study uses data collected during 11 Gaelic football games over the seasons 2019-2020-2021. During these games, players have worn two types of sensors: a 10 Hz GNSS sensor and a 100 Hz inertial measurement unit (IMU). The GNSS sensor captures positional data by recording latitude, longitude, and speed at 10 samples per second during play. Acceleration from GNSS data is computed as the rate of change of speed between consecutive time points. The IMU comprises a triaxial accelerometer and a triaxial gyroscope, each sampling at 100 Hz, providing 100 measurements per second along the x, y, and z axes.

5.2.2 Feature Engineering

Feature engineering is a critical step that transforms the raw, high-frequency GNSS and IMU sensor data into summarized descriptors of player activity. This process enables capturing complex workload patterns while reducing data dimensionality for effective machine learning modeling.

Per-Second Feature Extraction. The input raw sensor data were aggregated at 1-second frequency by computing the average for speed, acceleration, location (latitude and longitude), acceleration x, y, and z, and gyroscope x, y, and z in that second. This aggregation balances the need for temporal precision and computational manageability. Several key per-second features were computed, including:

- Metabolic Power: an estimate of instantaneous energetic demand calculated from sensor-derived metrics (Section 5.1.1). This variable is not directly provided by the GNSS or IMU sensors; therefore, it is computed using the formula described in Section 5.1.1. The metabolic power for each player in each second is computed and added as a new column to the original dataframe.

- Player Load and Gyroscope Load: measures of movement intensity variability and rotational dynamics computed from accelerometer and gyroscope signals, respectively [122, 123]. Player load is obtained by using the following formula:

$$\mathrm{PL} = \sqrt{(a_{x,t} - a_{x,t-1})^2 + (a_{y,t} - a_{y,t-1})^2 + (a_{z,t} - a_{z,t-1})^2} \tag{5.5}$$

  PL is calculated as a variation of the vector magnitude, determined by taking the square root of the sum of the squared differences in acceleration over time across the three movement axes (x, y, and z) [122].

  In the same way, we computed gyroscope load [123]. Gyroscope load can be described as how fast the direction of the gyroscope's rotation changes between one moment and the next. It is measured by combining the rotation speeds around all three axes into a single value that shows the overall rate of directional change [123].

- Distance Metrics: spatial features quantifying player positioning relative to teammates. Using the Haversine formula, three key spatial metrics were computed for each player-second combination: minimum distance to teammates, maximum distance to teammates, and average distance to teammates. These metrics provide insights into tactical positioning and collective movement patterns during gameplay.

- Cumulative Workload Features: progressive measures of accumulated high-speed movement, metabolic power, and acceleration counts capturing evolving player fatigue and exertion during games. Four progressive workload metrics, describing the running activity of players since the beginning of the game, were calculated: cumulative high speed distance (distance covered above 5.5 m/s speed threshold), acceleration count (number of accelerations exceeding 2 $\mathrm{m/s^2}$), cumulative metabolic power, and cumulative high metabolic load

distance (total distance covered above 20 W/kg threshold).

These per-second features and sensor-derived per-second variables were aggregated into 3-minute intervals using a sliding window approach.

3-Minute Window Aggregation and Feature Extraction. In this step, the shape of the dataset was modified by aggregating data into 3-minute time windows and extracting features describing the player's activity within each window. The choice of a 3-minute sliding window for data aggregation is supported by findings in elite soccer research, where moving averages of 1, 3, and 5 minutes have been used effectively to capture peak running performance and workload fluctuations [113, 124, 125], and in elite rugby union, where 3-minute epochs have been applied to quantify worst-case movement demands [126]. A 3-minute window was selected here as a balance between temporal granularity and data volume: it is short enough to capture meaningful short-term transitions in metabolic workload, which is the target of the predictive model, while providing sufficient data points per window for stable feature extraction.

The input dataset is composed of the sequences of variables returned by GNSS and IMU sensors, along with metabolic power, distance, and cumulative metrics describing, for each player, 1 second of activity during games. By using sliding 3-minute windows, we extract from these variables a set of features described in Table 5.1. The full set of features extracted at this stage and shown in Table 5.1, serves as inputs for subsequent predictive modelling.

5.2.3 Labelling

This section introduces the Metabolic Work Ratio (MWR), a derived metric expressing the energetic demand of player activity over fixed time periods, and explains how the workload labels for prediction are created.

Leave-1-Game-Out Cross-Validation. For each iteration of the analysis, one full game was extracted as the test set, with the remaining games used as the training set. Using the training data, player-specific MWR baselines were computed for label

Table 5.1: Features extracted from GNSS and IMU variables.

Variable	**Features Extracted**
Metabolic power	Average, Max, Cumulative metabolic power (total accumulated metabolic power from the start of the game up to the current window), High metabolic load distance (distance covered when metabolic power $\geq$ 20 $W \cdot kg^{-1}$), Cumulative high metabolic load distance, Percent time in zones (1: <10, 2: 10–20, 3: 20–35, 4: $\geq$35 $W \cdot kg^{-1}$), Trend (it fits a simple linear regression model to the sequence of values and returns the slope coefficient), Intensity drop (calculates the mean of the first and last quarters of the window, and returns the relative decrease as an index), Work-rest ratio (counts the total duration spent at high-intensity (¿ 20 $W \cdot kg^{-1}$) versus resting).
Speed	Average, Max, Percent time standing/walking ($\leq$2 m/s), jogging/running (2–5.5 m/s), high-speed running/sprinting (>5.5 m/s), Trend, Intensity drop, Distance covered, Cumulative distance
Acceleration	Average, Max, Number of accelerations (>2 m/s^2), Number of decelerations (<-2 m/s^2), Trend, Intensity drop, Mean acceleration jerk
Player Load	Average, Standard deviation, Count of peaks (>90th percentile)
Gyroscope Load	Average, Standard deviation, Count of peaks (>90th percentile)
Latitude & Longitude	Avg. distance to others, Avg. min/max distance to others, Betweennees centrality score (nodes are the players and edges the distances)

generation. Machine learning predictions were then performed on the test game using these labels. The process was repeated with a different test game, recalculating MWR baselines from the revised training set before the model retraining and testing. This iterative approach, called Leave-1-Game-Out Cross-Validation, ensures that model evaluation is performed on held-out games not seen during training, a strategy employed across sports analytics to assess model generalization [127], [128], [129]. It should be noted that strict temporal separation between training and test sets is only guaranteed in the fold where the chronologically last game is held out as the test set. In all other folds, some training games are temporally later than the test game, meaning future information is implicitly available during training. This is an inherent limitation of applying LOGO cross-validation to a small dataset of 11 games, where a fully chronological holdout scheme would leave insufficient training data. The approach is retained as the most practical evaluation strategy available given the dataset size, but results should be interpreted with this constraint in mind. Furthermore, since all data derive from a single team, the model may capture team-specific movement tendencies rather than generalisable workload patterns, which

limits external validity.

From a practical standpoint, LOGO cross-validation is computationally manageable for datasets of this size, as each fold trains on 10 games and tests on 1, requiring only 11 model fits per experiment. However, a known disadvantage is that with only 11 games, each training fold contains limited diversity in terms of match context, opposition quality, and seasonal variation, which may reduce the model's ability to generalize to unseen conditions. An alternative approach would be a fully chronological holdout, where only games occurring before the test game are used for training, eliminating any temporal leakage. This was not adopted here as it would leave the earliest games with very little training data, making model fitting unreliable. A further alternative is k-fold cross-validation stratified by game, though this does not respect the temporal structure of the data. LOGO was therefore retained as the most appropriate strategy given the constraints of the dataset.

Player-Specific Average Metabolic Power in the Training Set. The MWR is computed separately for each player because the average metabolic power in the training set ($\bar{MP}$), which is used to derive MWR, varies between players. For each player, the average metabolic power $\bar{MP}$ is calculated using all available data from the games in the training set. Since metabolic power values are recorded every second during these games, $\bar{MP}$ represents the average of all second-by-second values across the entire training set, as shown in Equation 5.6:

$$\bar{MP} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} MP_i, \tag{5.6}$$

where MP_i is the metabolic power at each second i, and N_{train} is the total number of seconds the player was active in the games of the training set.

State Segmentation. For each player and each 3-minute window in a given game, metabolic power values for each second ($MP(t)$) are classified as active or recovery states depending on whether the observed value is above or below the player's average metabolic power from the training set, as shown in Equation 5.7:

$$\text{Action State}(t) = \begin{cases} 1 & \text{if } MP(t) > \bar{MP}_{\text{player}} \\ 0 & \text{if } MP(t) \leq \bar{MP}_{\text{player}} \end{cases} \tag{5.7}$$

State changes are detected using cumulative shifts to identify transitions (e.g., active → recovery → active), which help segment continuous periods of activity or recovery.

Segment Area Calculation. In this step, the deviation of metabolic power from the player's baseline is quantified during each continuous active or recovery segment. This is done by calculating the area between the metabolic power curve and the baseline within each segment. This area provides a measure of elevated or reduced energy expenditure over time.

- Active Area (AA): sum of areas from all segments where $MP(t) > \bar{MP}(t)$, formally defined in Equation 5.8:

$$\text{AA} = \sum_{\text{active segments}} \int_{t_{\text{start}}}^{t_{\text{end}}} \left(MP(t) - \bar{MP}_{\text{player}} \right) \, dt \tag{5.8}$$

  The AA is the sum of the area under the curve formed by consecutive metabolic power values higher than the player-specific average metabolic power in the training set and the straight line of this average metabolic power. This area quantifies elevated energy expenditure.

- Recovery Area (RA): sum of absolute areas from all segments where $MP(t) \leq \bar{MP}(t)$, formally defined in Equation 5.9:

$$\text{RA} = \sum_{\text{recovery segments}} \left| \int_{t_{\text{start}}}^{t_{\text{end}}} \left(MP(t) - \bar{MP}_{\text{player}} \right) \, dt \right| \tag{5.9}$$

The RA is the sum of the area under the metabolic power curve during periods when the athlete's metabolic power is below their average metabolic power. This area represents periods at lower metabolic workloads, likely reflecting recovery or periods at less metabolic intensity.

MWR Computation. The formula to compute MWR for the window is expressed in Equation 5.10:

$$\text{MWR} = \frac{\text{AA}}{|\text{RA}| + \epsilon} \cdot t = \frac{\sum_{\text{active}} \int_{t_{\text{start}}}^{t_{\text{end}}} \left(MP(t) - \bar{MP}_{\text{player}}\right) dt}{\sum_{\text{recovery}} \left| \int_{t_{\text{start}}}^{t_{\text{end}}} \left(MP(t) - \bar{MP}_{\text{player}}\right) dt \right| + \epsilon} \cdot t, \tag{5.10}$$

where t is the size in minutes of the time window, ϵ a very small positive number to avoid division by 0 when no recovery occurs ($|\text{RA}| = 0$), AA and RA are the active and the recovery areas formally defined in Equation 5.8 and 5.9, respectively. The factor t, which is the length of the time window in seconds, adjusts the ratio so that the MWR reflects the total metabolic effort relative to the duration of the window. This scaling makes it possible to compare MWR values calculated from windows of different lengths.

In summary, MWR quantifies the ratio of cumulative metabolic demand above the player's baseline to cumulative demand below baseline within a fixed time window, scaled by the window duration.

Figure 5.2 shows the metabolic power profile of a randomly selected player over a short 30-second interval. The blue line traces the metabolic power $MP(t)$ recorded each second t, while the red dashed line represents the player-specific average metabolic power $\bar{MP}$ computed from their training data, serving as an individualized baseline. The green striped areas indicate periods when metabolic power exceeds the baseline (Active Area), reflecting bursts of elevated metabolic power output. The grey striped areas correspond to times when metabolic power is at or below the baseline (Recovery Area), representing phases of reduced energy demand or recovery. This plot shows how the metabolic power values are divided into ac-

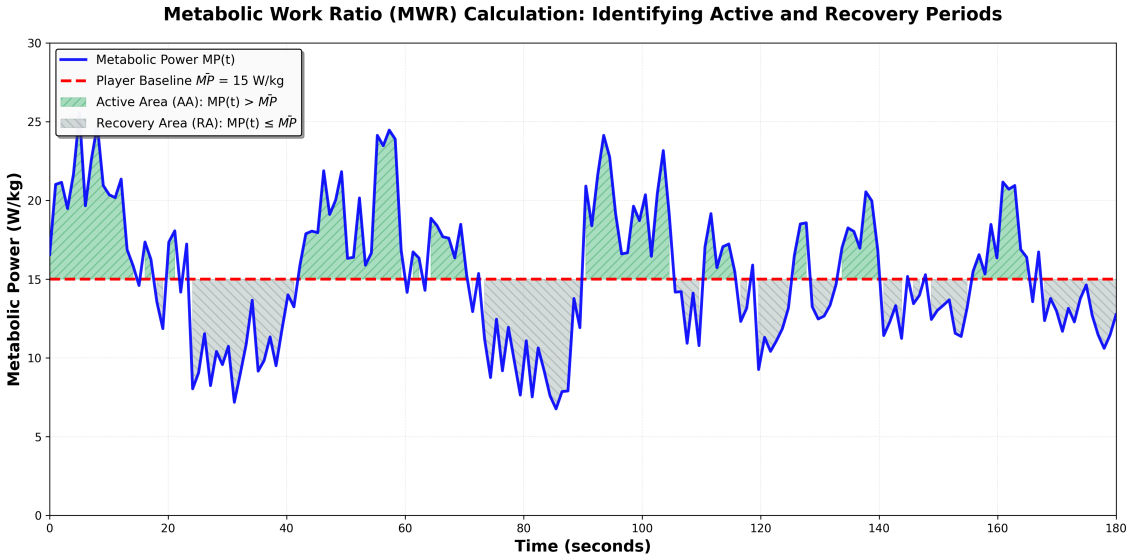


Figure 5.2: Active and Recovery Areas obtained from metabolic power trend and average metabolic power.

tive and recovery periods during a short time frame. Although the Metabolic Work Ratio (MWR) was calculated over 3-minute periods, displaying the entire 3 minutes would render the plot difficult to interpret. Therefore, only a 30-second segment is shown to clearly illustrate the separation and measurement of active and recovery areas

After computing MWR, MWR_{change} is then calculated. MWR_{change} quantifies the relative variation in a player's workload intensity between consecutive 3-minute windows. The metric is normalized to the player's historical average metabolic work ratio in the training set ($M\bar{W}R_{player}$), enabling individualized comparisons of workload dynamics. The formula is expressed as:

$$MWR_{change} = \frac{MWR_{t+1} - M\bar{W}R_{player}}{M\bar{W}R_{player}}, \tag{5.11}$$

where MWR_{t+1} is the metabolic work ratio calculated for the next 3-minute window, and $M\bar{W}R_{player}$ the player-specific average MWR computed across all training data. Our goal is to use features extracted from 3 minutes of data combined with cumulative features from the beginning of the game to predict the MWR_{change}

in the next 3-minute window.

MWR_{change} can be interpreted in the following way:

- Positive MWR_{change}: the athlete's MWR in the next 3-minute window is higher than their average MWR. This suggests an increase in the metabolic expenditure relative to their typical level. The higher the value, the larger the increase.

- Negative MWR_{change}: the athlete's MWR in the next 3-minute window is lower than their average MWR, indicating a decrease in the metabolic expenditure relative to their typical level. The more negative the value, the larger the decrease.

- MWR_{change} near to zero: the athlete's MWR in the next 3-minute window is close to their average MWR, suggesting a stable or consistent metabolic expenditure.

3rd Quantile Labeling. After computing MWR_{change} values for each 3-minute window in the training data, we calculate the player-specific threshold as the third quartile (75th percentile) of the training distribution of MWR_{change} values. Formally, for player p with training set T_p:

$$\theta_p = Q_3\left(\{MWR_{change}^{(t)} : t \in T_p\}\right),$$

where Q_3 denotes the third quartile function and t a time window taken from the training set.

We then assign binary labels $y_p^{(t)}$ for each time window t in both training and test sets as:

$$y_p^{(t)} = \begin{cases} 1, & \text{if } MWR_{change}^{(t)} > \theta_p \\ 0, & \text{otherwise} \end{cases}$$

This procedure transforms the continuous MWR_{change} values into a binary classification problem by identifying high workload transitions (label 1), which correspond to the top 25% of metabolic workload fluctuations for each player. These metabolic workload thresholds are player-specific and recalculated for each training fold, enabling adaptive and individualized labeling that reflects unique workload profiles.

This approach focuses the model on detecting the most critical workload changes likely related to increased physiological stress, rather than predicting all small variations.

Instead of relying on arbitrary or expert-defined thresholds (which may not generalize across players or sports), data mining techniques identify natural divisions in the data based on underlying patterns and variations. Automatically detected thresholds techniques adapt to unique workload profiles of players, ensuring that labels reflect individualised or team-based performance metrics rather than generic cut-offs [99]. Data mining approaches have been applied in sports science research [99], [130], [62], [37], with k-means clustering specifically used in several studies [130], [99],[62].

5.2.4 Model Pipeline

The model pipeline defines the process for training, validating, and interpreting machine learning models aimed at predicting player workload transitions. To ensure robust and realistic evaluation, a cross-validation strategy tailored to the sequential and game-structured nature of sports data is employed. The pipeline includes techniques to handle class imbalance in the target labels, comparison of multiple classification algorithms, and post-hoc interpretability analyses to explain model predictions.

The prediction task is designed as a binary classification problem. For each player and each 3-minute time window, the objective is to classify whether the metabolic workload change (MWR_{change}) in the subsequent window exceeds a player-specific

threshold, defined as the 3rd quartile of their historical MWR_{change} values, indicating a high workload transition (label 1), or remains below this threshold (label 0). This approach prioritizes detecting increases in energetic demand over predicting exact continuous values.

Leave-One-Game-Out Cross-Validation. To evaluate the predictive performance and the robustness of classifiers trained on event-level graph and action features, a *Leave-One-Game-Out (LOGO)* cross-validation scheme is adopted. LOGO cross-validation mimics a real-world scenario where data from past games are used to predict outcomes in the current game.

In this strategy, one game is used in turn as a testing set, while all remaining games form the training set. This approach ensures independence between training and testing data by preventing information leakage across game boundaries, which is critical in spatio-temporal sports datasets where observations within a game are temporally and contextually dependent.

Formally, given a dataset split into G distinct games:

$$\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\},$$

for each fold $g \in \{1, \ldots, G\}$, the training set is

$$\mathcal{D}^{(g)}_{\text{train}} = \bigcup_{\substack{j=1 \\ j \neq g}}^{G} \mathcal{G}_j,$$

and the test set is

$$\mathcal{D}^{(g)}_{\text{test}} = \mathcal{G}_g.$$

This ensures that the model evaluation simulates the realistic task of predicting MWR_{change} on entirely unseen games.

Within each fold, all preprocessing steps are conducted only on the training data to avoid data leakage. The trained model is then evaluated on the remaining test game, and performance metrics are recorded.

The calculation of MWR_{change} and the division into classes are repeated each time. Indeed, both the player-specific average MWR in the training set ($\bar{MWR}_{player}$), and the 3rd quantile of the distribution of MWR_{change}, change at each iteration as we change the composition of the training set.

Handling Class Imbalance. The target variable MWR_{change} (class 1) exhibits significant class imbalance, with fewer positive samples (MWR_{change} above the 3rd quantile) compared to negative ones. This imbalance represents a challenge for supervised learning algorithms, which may become biased toward predicting majority class outcomes, with the effect of underperforming on the minority class.

To address this problem, the *Synthetic Minority Oversampling Technique (SMOTE)* is used to balance the class distribution during training [131]. SMOTE works by generating artificial examples of the minority class through linear interpolation between existing samples in the feature space. Given a minority class sample $\mathbf{P}$ and one of its k nearest minority class neighbors $\mathbf{Q}$, a synthetic sample $\mathbf{Z}$ is generated as:

$$\mathbf{Z} = \mathbf{P} + u \cdot (\mathbf{Q} - \mathbf{P}), \quad u \sim \mathcal{U}(0, 1)$$

where $\mathcal{U}(0, 1)$ denotes the continuous uniform distribution [131]. This formulation places $\mathbf{Z}$ somewhere along the line segment between $\mathbf{P}$ and $\mathbf{Q}$.

SMOTE was applied only to the training folds during cross-validation to avoid any data leakage into the test data. During each *Leave-One-Game-Out* (LOGO) split, synthetic samples were generated within the training set after feature selection but before model fitting.

It should be acknowledged that standard SMOTE carries known limitations when applied to time-dependent data, as linear interpolation between samples from different time points may generate synthetic instances that do not respect the temporal ordering of the original sequence [132]. Alternatives such as SMOTE-ENN, which combines oversampling with edited nearest neighbour cleaning to clarify class boundaries [133], or SMOTE-TS, which is specifically designed to preserve temporal

dependencies during oversampling, would address this issue more directly. Anomaly detection approaches could also be considered, given that high MWR transitions represent extreme values by definition [134]. Standard SMOTE was used here because the features used for the classifier are window-level aggregates rather than raw time series: each sample represents a summarised 3-minute window, and the temporal ordering of windows is not directly encoded in the feature vectors, which reduces but does not remove the concern about temporal interpolation. Future work should evaluate whether SMOTE-TS or anomaly detection methods improve classification performance in this setting. Regarding causality, player substitutions and red cards represent a known issue: when a player leaves the game, the absence of their data may cause abrupt transitions in the MWR distribution that do not reflect genuine workload changes. This was partially addressed by the contextual filtering described in Section 3.2.1, though some cases may remain.

Machine Learning Models and Feature Set. Our analysis is structured around two main experiments, conducted sequentially to answer distinct research questions.

1. *Experiment 1: Feature Set Evaluation.* The first experiment employs the Gradient Boosting Machine (GBM) to assess the capability of different feature sets in predicting whether the Metabolic Work Ratio change exceeds a player-specific threshold. We compare three feature sets:

    - *Standard*: uses the full set of features described in Table 5.1, extracted from the current 3-minute window, plus cumulative features tracking accumulated workload since the start of the match. These features were chosen to capture the player's immediate physical state (current speed, metabolic power, acceleration patterns, player load, and spatial positioning) alongside their overall fatigue trajectory through cumulative metrics.
      .
    - *Lag 1*: augments the "Standard" feature set with the same features from the immediately preceding 3-minute window. The rationale is that a

player's workload in the next window may depend not only on their current state but also on what they were doing in the previous window, for example whether they are recovering from a recent high-intensity phase.

- *Lag 2*: augments the "Standard" feature set with features from the two preceding windows, extending the temporal context to 9 minutes of prior activity. This tests whether a longer recent history improves prediction, for example by capturing fatigue accumulation over a sustained period of play.

The objective is to compare predictive results across varying time windows to determine the optimal temporal horizon for achieving high accuracy. This analysis aims to understand how a player's recent history affects changes in their metabolic power over time.

2. *Experiment 2: Model Comparison.* Following identification of the best-performing feature set from Experiment 1, this set is used to evaluate the predictive performance of multiple machine learning models (Logistic Regression, Random Forest, Gradient Boosting, and XGBoost) for binary classification of metabolic workload transitions. This experiment evaluates which model architecture best captures metabolic workload dynamics on the selected feature representation.

In summary, both experiments address the same binary classification task of predicting metabolic workload transitions. However, in Experiment 1, the machine learning model (Gradient Boosting Machine) is fixed, and the focus is on identifying the best feature set by comparing different temporal contexts. In Experiment 2, the best-performing feature set from Experiment 1 is fixed, and the predictive performance of various machine learning models is evaluated to determine the best-performing model.

Post-Hoc Explainability via SHAP Analysis. SHAP (SHapley Additive exPlanations), a method for quantifying the contribution of each input feature to the model prediction, was employed to interpret the predictive models [135]. SHAP

is based on cooperative game theory, assigning each feature a Shapley value based on its marginal contribution to the model prediction.

Given a trained model $\hat{f}$ and an input instance $\mathbf{x} = (x_1, x_2, \ldots, x_M)$ with M features, SHAP constructs an explanation model g expressed as an additive feature attribution [135]:

$$g(\mathbf{z}') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j,$$

where:

- $\mathbf{z}' \in \{0,1\}^M$ is a simplified binary vector indicating the presence (1) or absence (0) of each feature in the coalition,
- $\phi_0 = \mathbb{E}_{\mathbf{x}}[\hat{f}(\mathbf{x})]$ is the expected value of the model output over the background dataset,
- ϕ_j is the SHAP value for feature j, representing its contribution to the deviation of the prediction $\hat{f}(\mathbf{x})$ from the expected output.

The SHAP value ϕ_j for feature j is defined as the weighted average of feature j's marginal contributions over all possible subsets S of features not including j [135]:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \Big[\hat{f}_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) - \hat{f}_S(\mathbf{x}_S) \Big],$$

where:

- $F = \{1, 2, \ldots, M\}$ is the set of all features.
- S is a subset of features excluding j.
- $|S|$ is the number of elements of the subset S.
- The factorial terms $|S|!$ and $(M - |S| - 1)!$ together with the denominator $M!$ constitute a weighting factor:

$$\frac{|S|!\,(M - |S| - 1)!}{M!}$$

This weighting corresponds to the proportion of all possible permutations of the feature set where the subset S appears before feature j.

- $\hat{f}_S(\mathbf{x}_S)$ is the model output when only features in S are present.
- $\hat{f}_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}})$ is the model prediction when the feature j is added to the subset S.
- The difference

$$\hat{f}_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}}) - \hat{f}_S(\mathbf{x}_S)$$

represents the *marginal contribution* of feature j to the prediction when added to the coalition S.

In our experiments, SHAP analysis was applied post-hoc to the best-performing classifiers. For each game, SHAP values for each feature were computed using the Python library SHAP [136]. Then, these values were averaged to obtain an importance score for each feature, along with the direction of its influence. This approach provides clear explanations of how different features and activity metrics affect metabolic power trends through the MWR framework and player-specific adaptive thresholds.

The SHAP values are averaged across all 11 games rather than computed for a single game. Since MWR is a physical metric grounded in consistent physiological principles, the relative importance of features is expected to remain broadly stable across games: the same physical signals that drive workload transitions in one match should drive them in another. Averaging across games reduces game-specific noise and provides a more stable and generalizable picture of feature importance than relying on a single game, which could reflect specific match conditions.

5.3 Validation and Discussion

Figure 5.3 shows, for each player except the goalkeeper, information regarding player participation in the sampled games. Players who played less than three games

were excluded to ensure reliability and consistency when interpreting results. These players represent a small sample size, often with highly variable or very low playing times, which can introduce noise. In addition, for the test set, instances related to players who played less than five minutes were removed. Such brief appearances may not provide sufficient data to model player performance effectively.

Beyond participation-related noise, several additional sources of variability should be acknowledged. First, GPS and IMU sensor accuracy introduces measurement uncertainty, particularly at high velocities and during non-linear movements, which may affect the precision of metabolic power estimates derived from positional data [46], [96]. Second, contextual factors inherent to match play (opposition quality, scoreline, tactical formation, pitch dimensions, and weather conditions) are not captured in the sensor data and may substantially influence the physical demands placed on individual players within any given 3-minute window. Third, positional role changes, mid-game tactical substitutions, and periods of reduced involvement due to game state may create within-game variability that the model cannot distinguish from genuine workload transitions. These limitations mean that prediction errors in some games and for some players may reflect contextual variability rather than model inadequacy and should be considered when interpreting the results.

For each experiment, one game is iteratively selected as testing set while all the remaining as training set, ensuring the LOGO cross-validation scheme defined in Section 5.2.4. Following the methodology described in Section 5.2, the MWR_{change} is computed for each player, and a binary label is assigned based on the player-specific threshold derived from the training distribution. The machine learning model is then trained on the training set and evaluated on the test game. This process is repeated for every game in the dataset to ensure robust and generalizable evaluation of model performance across different match conditions.

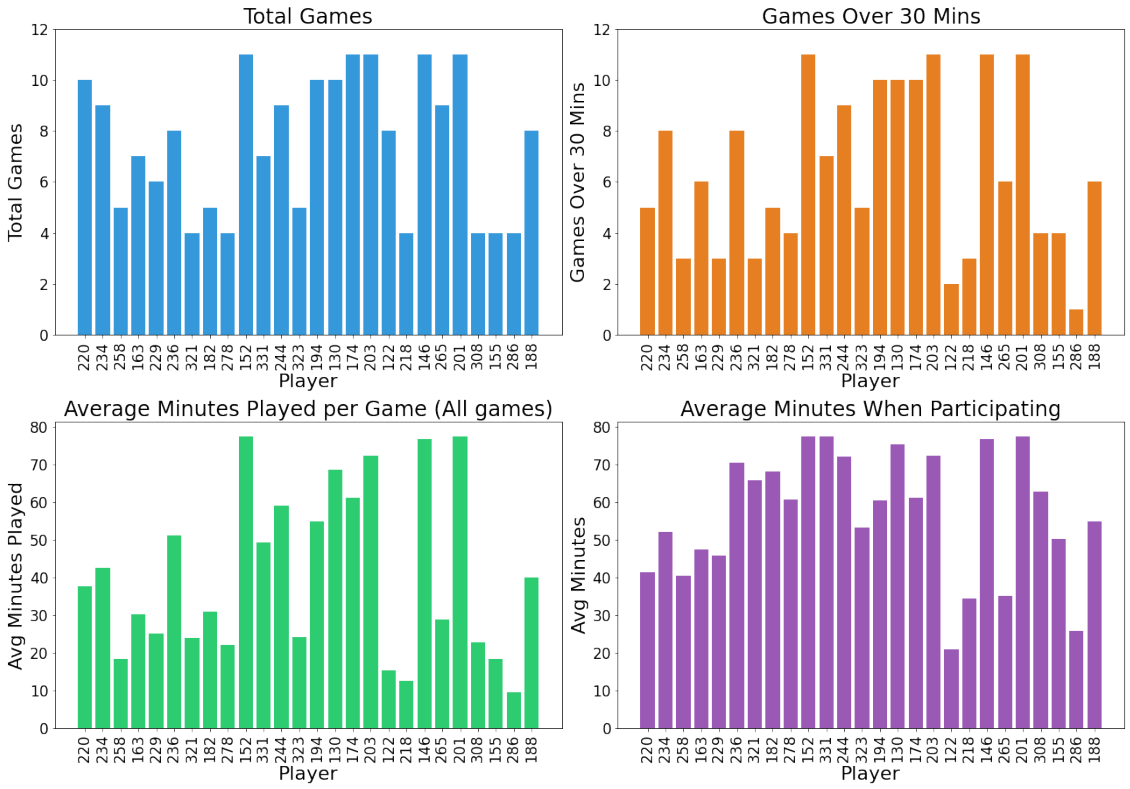


Figure 5.3: This figure compares each player's total games, games with over 30 minutes played, average minutes per game, and average minutes when participating, highlighting variations in participation and playing time across the team.

5.3.1 Experiment 1: Feature Set Evaluation

The first experiment uses Gradient Boosting Machine to evaluate the predictive performance of the model trained on different feature sets, comparing the "Standard" feature set with lagged feature sets ("Lag 1" and "Lag 2") that incorporate data from previous time windows. This comparison aims to determine whether including temporal context from prior windows improves the classification of metabolic workload transitions. The binary classification task remains consistent across sets, with models predicting whether the metabolic work ratio change exceeds the player-specific threshold described earlier.

Table 5.2 summarizes the average accuracy and F1 scores achieved by GBM when trained on different feature sets. The F1 score is first calculated for each test game in the LOGO cross-validation, and then the average and standard deviation across all games are reported to summarize overall performance. The "Standard" feature

Table 5.2: Comparison of predictive performance across different feature sets.

Features Set	**Avg F1 Score**	**Avg Accuracy**
Standard	0.69 ± 0.05	0.71 ± 0.04
Lag 1	0.68 ± 0.06	0.70 ± 0.06
Lag 2	0.68 ± 0.07	0.71 ± 0.05

set includes metrics extracted from the current time window and cumulative values since the start of the game, while "Lag 1" and "Lag 2" feature sets augment the "Standard" features with data from one and two previous time windows, respectively (Section 6.2.4). The results indicate that adding information from previous time windows does not increase the performance of the machine learning model: the F1 score for the "Standard" feature sets across the 11 games is the highest and the most stable (0.69 ± 0.05). The average accuracy does not vary significantly for the diverse feature set (0.71, 0.70, and 0.71), with "Standard" features obtaining more stable results (standard deviation equal to 0.04, 0.06 and 0.05, respectively). The analysis demonstrates that prediction of metabolic power trends for the next 3 minutes is most effectively achieved using the player's activity data from the current 3-minute period, without increasing the temporal window. Some temporal information is still included in the cumulative features collected on some specific variables (Section 5.2.2), which return a history of the evolution of these features since the beginning of the game.

5.3.2 Experiment 2: Model Comparison

A comparison of the results of machine learning models using the "Standard" feature set individually for each tested game is shown in Figure 5.4. The classification results across 11 test games highlight the effectiveness of tree-based ensemble methods for metabolic workload pattern recognition. Gradient Boosting (GB) achieves the highest F1 scores in 6 of 11 games, while Random Forest (RF) in 4 games. The superior performance of GB and RF highlights the importance of hierarchical feature interactions and non-linear decision boundaries in predicting metabolic workload transitions. However, model efficacy varies substantially across teams: for example,

GB's F1 drops to 0.58 in Game 997 but peaks at 0.80 in Game 947, implying game-specific factors influence predictability. XGBoost shows intermediate performance. Overall, GB emerges as the most robust model, balancing predictive power (average F1: 0.67 ± 0.06) and consistency across datasets.

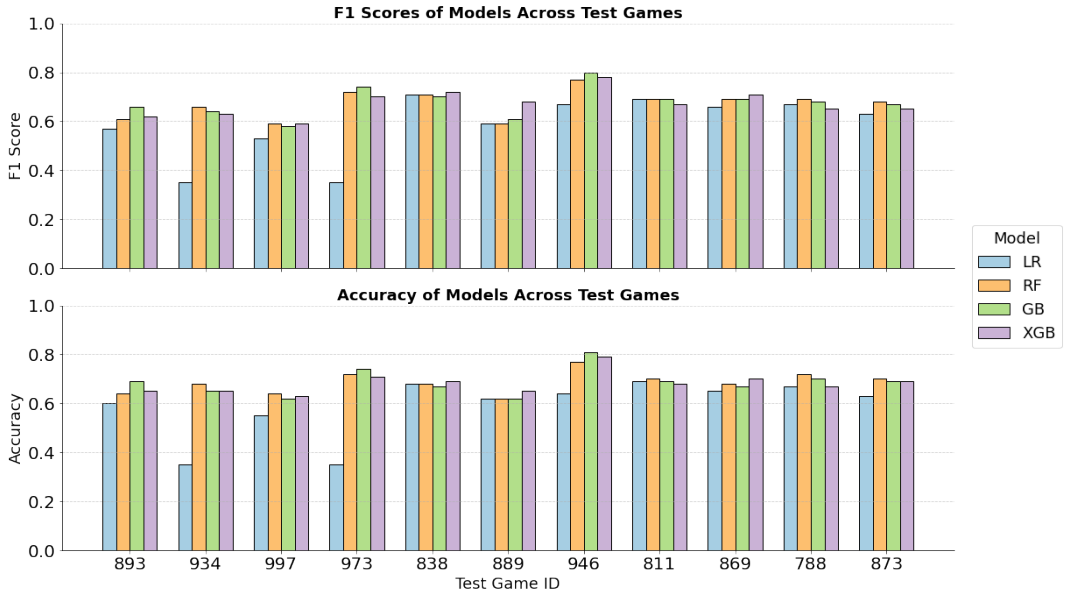


Figure 5.4: This figure presents a comparative analysis of Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), across multiple test games, showing their F1 scores and accuracy to evaluate their predictive performance.

Figure 5.5 illustrates the conditional distribution of subsequent MWR classes based on the MWR class in the current window, offering insight into the temporal stability and transition dynamics of workload states. When the current MWR class is 0, a majority (77.5%) of subsequent windows remain in class 0, while 22.5% transition to class 1. In the same way, when the current class is 1, 69.2% of the next windows diverge to class 0, with 30.8% persisting to class 1. This pattern indicates moderate persistence within each workload class, with a higher likelihood of remaining in a lower workload state (class 0) compared to a higher workload state (class 1). The increased probability of transitioning from class 0 to class 1 (30.8%) suggests that high workload periods are often followed by a stability in intensity, potentially reflecting intensive game moments. These results are highly dependent on the chosen size of the time window, 3 minutes in this study.

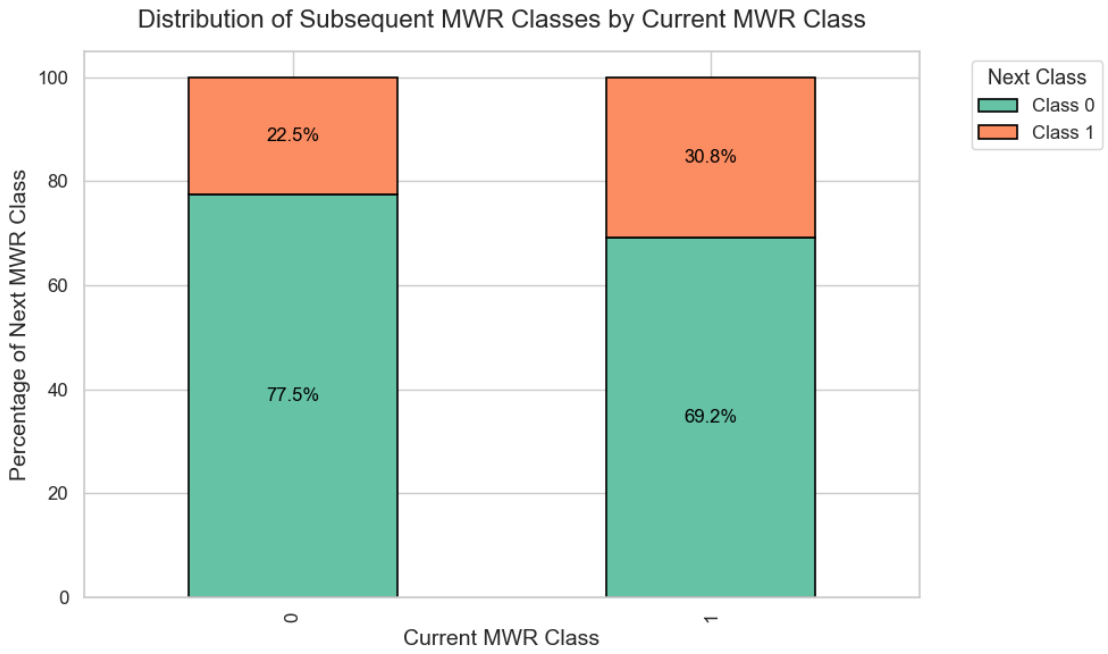


Figure 5.5: Count of classes in the next time window based on the current class.

The boxplot in Figure 5.6 shows the distribution of F1 scores, obtained using the GMB model, across players, with colors indicating the number of games played. In general, the model tends to obtain better results for players with a higher presence in games (yellow and green colors), with some exceptions. For example, Players 220 and 194, despite their elevated presence, exhibited considerable variability in F1 scores. On the contrary, Players 174, 146, and 201 combined high involvement with stable and elevated F1 performance. Some players with lower levels of participation, such as Players 323 and 308, demonstrated consistently high F1 scores. The model reaches lower results with players with less presence (example: Players 163, 122, 218, 155, and 286).

In summary, the model's performance is more robust for those players with more participation, reflecting their higher presence both in the test and in the training set. This may indicate that the model learns patterns of player behavior when it is exposed to a sufficient number of relevant personal examples. This is confirmed by the low average F1 score those players who played a lower number of games.

Figure 5.7 illustrates the distribution of the GBM model's weighted F1 scores across consecutive time windows in the first and second halves. In the first half,

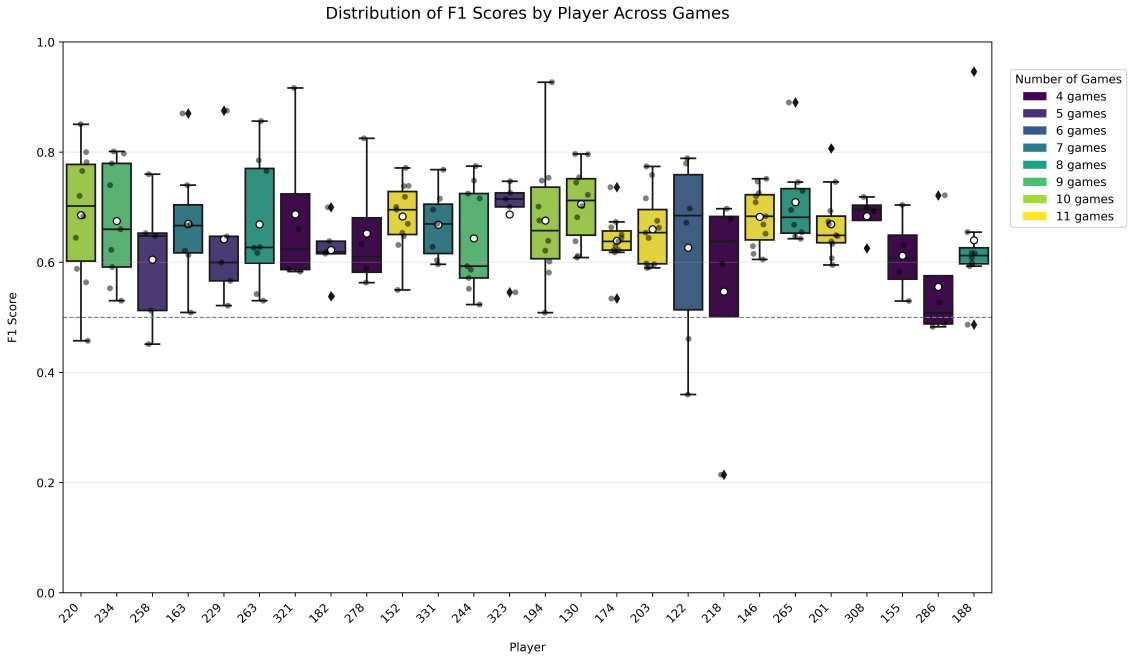


Figure 5.6: Boxplots showing the distribution of F1 scores per player across games, with colors indicating the number of games played and individual points representing single-game performances.

the F1 score starts at a moderate level and exhibits a generally upward trend, with some oscillations and local peaks, such as for the 540–720 and 1440–1620 windows. The model's performance stays relatively stable and finishes the half with one of its highest values. The second half displays greater variability and more peaks, with F1 score reaching the highest and lowest values. After a drop in the middle of the half, the model's performance follows a growing trend towards the end of the half, with a drop in the last window. The model demonstrates robust predictive performance throughout the match, with a tendency for higher accuracy in the later stages of both halves. Interestingly, both halves follow the same pattern in the central part, with F1 scores reaching the same score in two consecutive windows (540-720) and (720-900).

5.3.3 Insights on Features

The SHAP summary plot in Figure 5.8 illustrates the mean feature contributions to the prediction of Class 1 (high MWR_{change}) across all games obtained using GB in combination with "Standard" feature set. Cumulative metabolic power (Cumula-

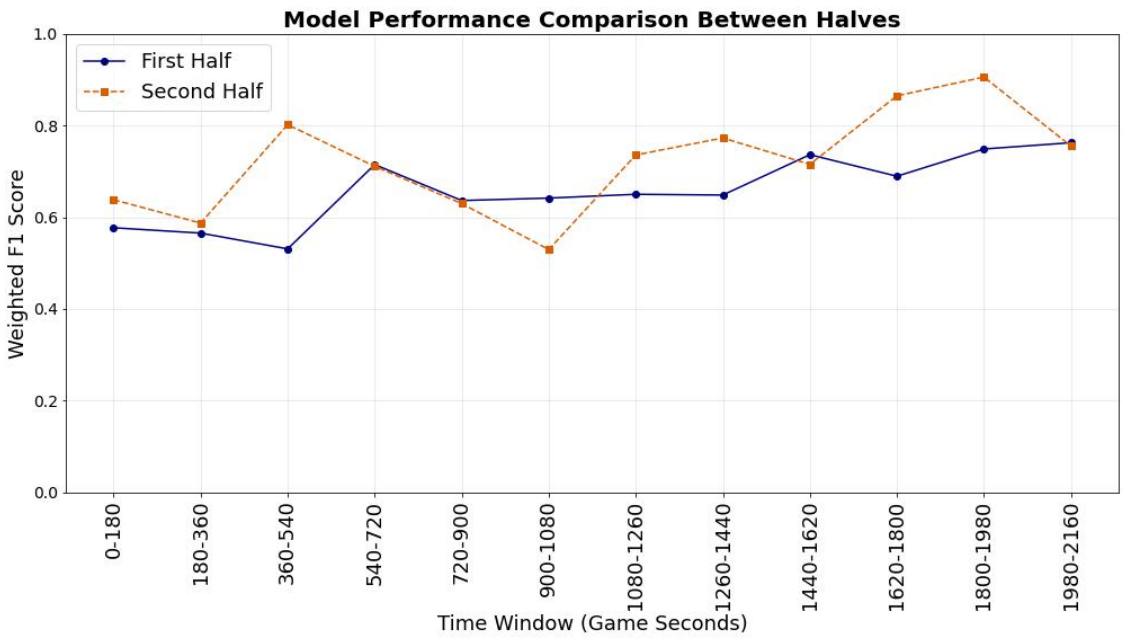


Figure 5.7: F1 score of the predictive model across sequential time windows for the first and second half of the matches.

tive_mp), cumulative distance (cumulative_distance), and cumulative high-metabolic load distance (cumulative_hmld) are the most influential positive predictors, suggesting that higher values of these features increase the probability of a high MWR_{change} event in the next window. This may indicate that the level of activity increases in later stages of the game. Percentage time in metabolic power zone 4 (Percent_time_zone4_mp) has the strongest negative influence, indicating that higher time spent in the most intense metabolic power zones is associated with a reduced likelihood of experiencing a transition to a high metabolic workload change (Class 1) event in the next window. Same situation for max speed (max_speed), suggesting how higher efforts in the current window are followed by recovery phases in the following window. Other features such as high-speed running or sprinting time in the window (hsr or sprinting), number of accelerations (num_accelerations), and standard deviation of player load values in the window (std_player_load), also show meaningful negative contributions, whereas metrics like average distance to teammates (avg_avg_dist), average minimum distance to teammates (avg_min_dist) and average maximum distance to teammates (avg_max_dist) positively influence the model's prediction of class 1. The mix of positive and negative SHAP values among top fea-

tures highlights the interplay between acute high-intensity exposure (which appears indicative of normal activity states) and cumulative workload (which elevates risk for significant workload changes). This pattern underscores how metrics on the current window and historical metrics contribute significantly in predicting transitions in athlete load profiles.

At the lower end of the influence spectrum, the five features with the smallest mean SHAP values were mean player load (mean_player_load), maximum metabolic power (max_metabolic_power), gyroscope load peak count (gyro_load_peaks_count), high metabolic load distance (hml_distance), and mean gyroscope load (mean_gyro_load). Four of these five features exert a negative influence on Class 1 predictions, while high metabolic load distance shows a weak positive contribution. The minimal influence of gyroscope-derived features suggests that rotational movement patterns at the 3-minute aggregation level do not meaningfully discriminate between workload transition states in this dataset, possibly because their effects are averaged out over the window duration. The low contribution of maximum metabolic power, despite its theoretical relevance to workload intensity, may reflect redundancy with the cumulative metabolic power feature which captures a similar physiological signal over a longer time horizon and emerged as the strongest positive predictor.

It should also be noted that SHAP values would be expected to vary during a game: in early windows, current-window intensity features are likely more informative, while cumulative features become increasingly dominant as fatigue accumulates in the following stages of the match. Computing SHAP values separately for early, mid, and late match windows to examine this within-game evolution is identified as a direction for future work.

By leveraging a rich feature set and machine learning models, our approach achieved high accuracy in forecasting transitions between high-intensity and recovery phases, as demonstrated in Figure 5.4, 5.6, 5.7. This demonstrates the potential for coaches to monitor player workload and make data-driven decisions based on interpretable indicators that reflect the demands of match play, provided the required

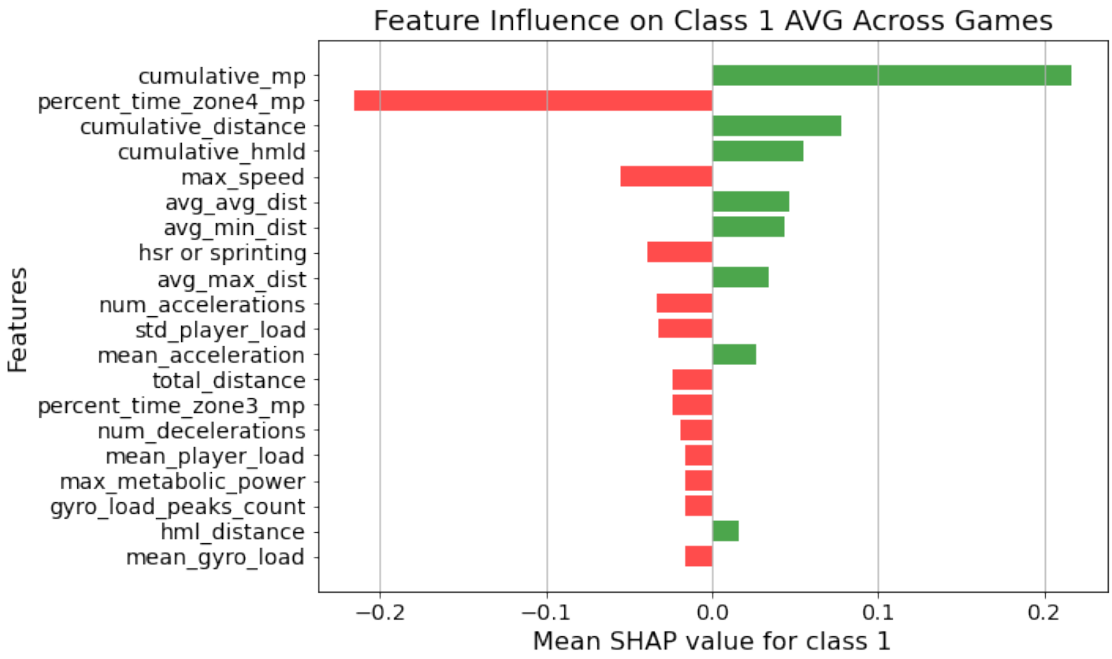


Figure 5.8: Direction of influence of the features to predict class 1 obtained by averaging shapley values across the tested games.

raw sensor data streaming infrastructure is available.

5.4 Summary

This work presents a new methodology to forecast the intensity of activity in elite Gaelic football players using advanced sensor data and machine learning techniques. A novel metric, the Metabolic Work Ratio (MWR), is introduced, which quantifies the balance between high-intensity work and recovery within a short time window based on metabolic power calculation. Data collected from 11 official matches using GNSS and IMU sensors were processed to extract features describing players' activity levels. These features were then reduced to train machine learning models for predicting the MWR in the subsequent 3-minute window. Our results indicate that the proposed method predicts with good accuracy whether a player's metabolic workload will significantly increase within the next 3-minute window, effectively capturing changes in the intensity of their physical activity.

While this study introduces a valuable framework for predicting transitions in

metabolic power states in elite Gaelic football, it is important to note that the results are influenced by the choice of time window used for feature extraction and prediction, which may affect the detection of brief changes in intensity. In addition, although the MWR offers a novel perspective, its physiological relevance requires further validation. Implementing this method during live matches would require continuous streaming of raw sensor data, which is not currently available through standard commercial monitoring platforms such as STATSports, where live dashboards display only aggregate summary statistics. This work therefore operates as a proof-of-concept, demonstrating that accurate prediction of metabolic workload transitions is achievable when raw data is accessible offline. Real-time deployment remains a direction for future work, contingent on advances in live raw data streaming infrastructure.

Chapter 6

Detecting Shot Events in Game Phases

The work presented in this chapter has been published in:

- Sheridan Dermot, Valerio Antonini, Michael Scriney, and Mark Roantree. "Automated Detection of Shot Events in Game Phases Using GNSS Data from a Single Team." In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, pp. 95–105. Cham: Springer Nature Switzerland, 2024. [52]

The first and second authors contributed equally to all aspects of this work, including: conception and design of the feature engineering methodology, implementation of the pairwise adjacency matrix approach, development and evaluation of the machine learning pipeline, analysis and interpretation of results, and drafting of the manuscript.

6.1 Introduction

Chapter 5 presented a machine learning framework for forecasting periods of high energy demand in elite Gaelic football players using wearable sensor data. The study predicted metabolic workload transitions over 3-minute windows, achieving robust

performance with Gradient Boosting model (average F1 score: 0.69 ± 0.05) using metrics extracted from GNSS and IMU sensor data.

In contrast to the previous chapters, this chapter addresses a distinct but related challenge in sports performance analysis: the creation and availability of contextual game data. While Chapters 3, 4, and 5 demonstrated various analytical approaches using sensor data, they assumed that contextual data would be available when understanding the context is needed. However, this assumption does not reflect the practical reality faced by many clubs and teams.

In professional or elite-level sports, contextual game data (referred to as log data in this research) is generated through manual video analysis. Sport analysts re-watch recorded matches and use specialized software to code specific events such as ball possession changes, attacking phases, defensive transitions, kickouts, and shot attempts with precise timestamps [137]. This manual annotation process provides information that can enrich the analysis of physical performance patterns obtained with sensor data. However, collecting log data in this way presents several limitations:

- *Time-intensive*: A single match requires hours of manual coding by trained sport analysts [137].
- *Expensive*: The cost of employing specialized analysts and software represents a significant resource commitment.
- *Prone to errors*: The coding process can be inconsistent and prone to errors even with standardized protocols [137].
- *Not scalable*: The manual approach cannot scale to analyze multiple matches simultaneously or provide real-time insights during live matches.
- *Retrospective gap*: Many clubs accumulated years of sensor data from wearable devices, but not the corresponding log data.

This creates a practical problem: while wearable sensor technology has become increasingly accessible and affordable, the ability to extract contextual insights from

this data remains limited to organizations that can afford dedicated video analysis resources. In addition, log data can sometimes report errors or inconsistencies. As a result, a large amount of collected sensor data remains underutilized, analyzed only for physical purposes without tactical interpretation [46], [108], [21].

Chapter 5 found movement patterns in how teams use different parts of the playing area, but without game logs, these patterns could not be linked to specific phases of play. In Chapter 5, predictive models could forecast changes in physical intensity, but without context, they could not explain which game situations caused those changes. If contextual information could be automatically extracted from sensor data, it would be possible to explain movement patterns found in Chapter 4 or changes in metabolic power observed in Chapter 5 without needing time-consuming manual video coding.

Chapter 7 addresses this challenge by investigating to what degree key game phases can be automatically identified using only sensor data, without requiring manually annotated log data. Specifically, machine learning models are investigated to determine if they can identify game phases (possession sequences from ball acquisition to loss of possession or shot attempt) directly from one team's movement patterns captured by wearable sensors.

6.1.1 Log Data: Collection and Structure

Log data are generated from video analysis, and teams use these data in combination with sensor data to enhance their performance by analysing both their and the opponents' gameplay systematically [138]. Log data are event-based: they are only recorded when a specific situation of play occurs in the game. Game footage from internal team video recordings and external media broadcasters was imported and coded using tagging software. All games were coded using Nacsport Scout Plus software (Scout Plus V 6.5.0, New Assistant for Coach Sport, S.L).

A trained sport analyst, employed by the team, re-watched each game and annotated specific playing events. Each recorded entry in the log data contains three

temporal components and a text description (plus some additional information not used in this analysis):

- *Start*: The timestamp marking the beginning of the playing event.
- *Click*: The timestamp marking the precise moment when the event actually happens. It represents the instant the analyst clicks to mark the key moment. For example, the exact moment a tackle is made, a shot is taken, or possession changes. In some cases, start and click can be the same timestamp.
- *End*: The timestamp marking the conclusion of the event. For example, when the ball goes out of play after a shot, or when possession is fully established by the opposing team.
- *Category*: A categorical label provided by the sports analyst describing the type of event, for example: 'TA Shot', 'TB Kickout', 'Turnover TA', 'Throw in'.

The log data contains various categorical labels that describe playing events during games, such as 'Throw in', 'Attack Team A', 'Attack Team B', 'TB Shot', 'TA Shot', 'TA Kickout', 'TB Kickout', 'TA Possession from KO', 'TB Possession from KO', 'Turnover TA', 'Turnover TB', and others. Here, 'TA' is the shortcut for Team A, while 'TB' for Team B. Throughout this chapter, Team A represents the team from which the sensor data were collected, while Team B represents the opposing team. Since the sensor data were collected on the same team, Team A remains consistent across all games.

An example of log data for a game is shown in Figure 6.1.

6.1.2 Selecting Shots Phases as Classification Target

In this research, *game phases* are defined as time intervals of ball possession that start when a team gains control of the ball and end when the same team loses possession or attempts a shot. The variety of categorical labels in log data necessitates

Category	Start	Click	End
TB Pos from Throw in-up	00:05	00:10	00:15
Throw in	00:05	00:10	00:15
Team A Pos frm TO	00:00	00:15	01:15
Turnover TB	00:15	00:15	00:44
Attack Team A	00:39	00:44	01:05
TA Shot(P)	00:50	00:50	01:05
TB Kickout	01:13	01:13	01:22
TA Pos frm KO	01:12	01:22	01:38
Attack Team A	01:18	01:23	01:38
TA Shot(P)	01:27	01:27	01:38
TB Kickout	01:49	01:49	01:57
TA Pos frm KO	01:47	01:57	02:14
TB Foul	01:55	02:00	02:05
Attack Team A	01:57	02:02	02:14
TA Shot(P)	02:07	02:07	02:14
Team B Pos frm TO	01:59	02:14	02:59

Figure 6.1: Example of log data for a game, showing sequential entries with start, click, and end timestamps, along with the relative categories provided by the sports analyst.

determination of which game phases should be prioritized for automatic classification. Figure 6.2 presents the frequency distribution of all categories across our dataset.

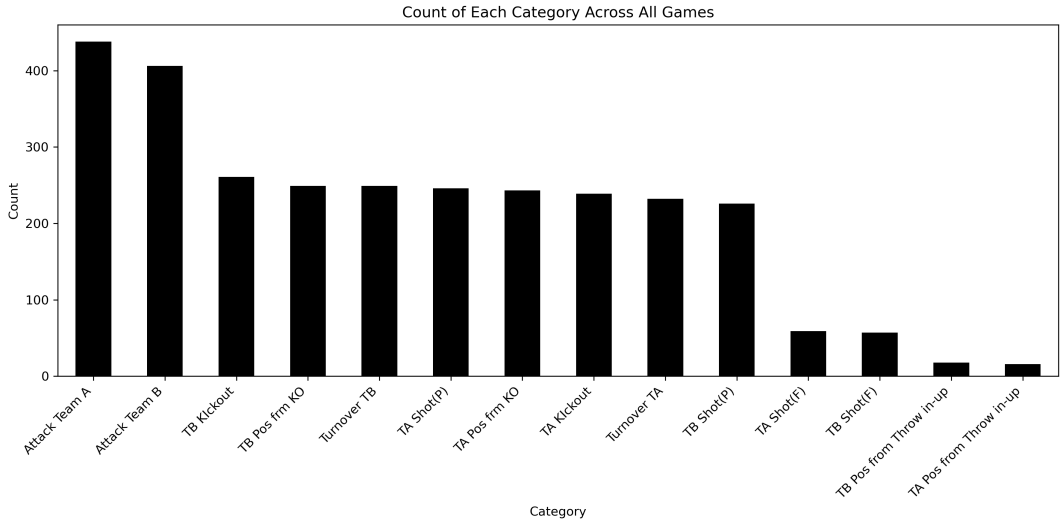


Figure 6.2: Distribution of category frequencies across all games.

Due to the different frequencies of each playing category shown in Figure 6.2, and the nature of the game, classifying all categories is challenging. Therefore, this chapter focuses exclusively on classifying, in two separate experiments, shot phases done by the two teams, represented by the categories 'TA Shot(P)' and 'TB Shot(P)'. Here, (P) denotes shots made during play (P) as opposed to shots made as free-kicks after a foul (F). This focus is motivated by:

- *Tactical significance.* Shots represent key moments that directly determine scoring opportunities [138] [38]. Thus, identifying when shooting phases occur provides coaches with critical contextual information about attacking and defensive situations.

- *Distinctive movement patterns.* Among all the situations of play, phases ending in shots are hypothesized to produce the most distinct sensor movement patterns in terms of speed, accelerations, and trajectories, consistent with findings in IMU-based event detection research in field sports [139], making them a practical starting point for automatic classification.

By demonstrating the feasibility of automatically detecting phases ending in shots using sensor data, this work could be extended to address remaining categories.

6.1.3 Experimental Design

Two different classification experiments using the same methodology are performed in this chapter:

- **Team A shot phases (Experiment 1)**: Game phases that end with a shot attempt by the team whose sensor data is being analyzed (Team A). These phases represent offensive sequences where the tracked team maintains possession and takes a shot. In our labelling scheme, phases ending with a 'TA Shot' event are assigned label 1, while all other phases are assigned label 0. The objective is to classify, using only Team A sensor data, when a game phase will end in a shot made by Team A.

- **Team B shot phases (Experiment 2)**: Game phases that end with a shot made by the opposing team (Team B). When Team B gains possession, the game phase begins, and if it concludes with a 'TB Shot' event, it is labelled with 1. All the remaining game phases are labelled with 0. The objective is to classify, using only Team A sensor data, whether the opponent team (Team B) took a shot.

The primary objective of this chapter is to explore the feasibility of automatically classifying shooting phases based on sensor data from only one team, without relying on manually coded log data or data from the opposing team. While existing log data are used to label game phases, the goal is to develop models that can operate only with sensor data from one team.

Existing approaches to automatic game phase and event detection in team sports rely on optical tracking systems providing full positional data for both teams and the ball, using ball-centric features such as ball velocity, acceleration, and proximity to players to classify events [90], [91], [92], [93]. These methods achieve high detection accuracy but require expensive infrastructure unavailable to most clubs. Where wearable sensor data has been used for event detection, studies have focused on individual player actions using ball-contact IMU signals rather than collective team movement patterns [139]. Relational representations of player interactions have been explored in graph-based tactical analysis [68], [70], but have not been applied to game phase classification from wearable GPS and IMU data. Critically, no prior work has addressed automatic game phase detection using only single-team wearable sensor data without ball position or opponent tracking data. This chapter addresses this gap through a novel feature engineering methodology that constructs adjacency matrices from pairwise differences in spatial and inertial metrics (distance, speed, 3D acceleration, and 3D gyroscope) between every pair of players during each game phase. These matrices encode the collective inter-player coordination patterns of a single team and serve as input features for supervised machine learning models, representing a methodological departure from both ball-centric event detection and individual player activity recognition approaches.

6.2 Methodology

This section describes the methodology to automatically classify phases ending in shots using sensor data from a single team. The methodology is composed of the following sequential steps: extraction of game phases from log data, feature extraction

from sensor data, feature selection, and model building.

Figure 6.3 provides a visual overview of the full methodology pipeline described in this section.

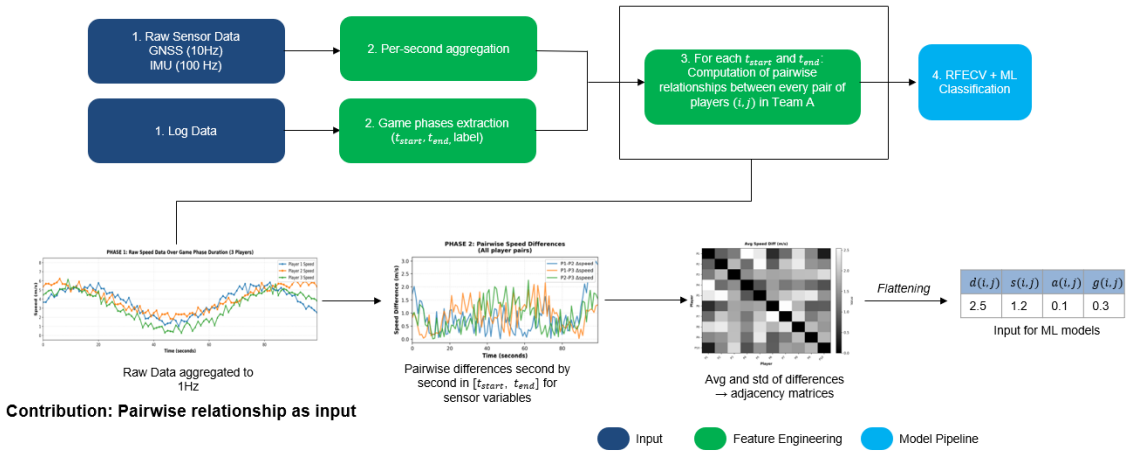


Figure 6.3: Overview of the methodology for shot event detection. Raw GNSS and IMU sensor data and log data are processed through pairwise feature engineering to construct adjacency matrices, which are flattened into feature vectors and fed into machine learning classifiers following RFECV-based feature selection.

6.2.1 Game Phases Extraction from Log Data

The first step involves identifying game phases from the log data and extracting the timestamps that mark their start and end. In Section 6.1.2, game phases are defined as time intervals of ball possession that start when a team gains control of the ball and end when the same team loses possession or attempts a shot.

A team gains possession after one of the following events:

- The start of a half, after a throw-in.
- A turnover, including situations where the team in possession commits a foul, resulting in possession transferring to the opposing team.
- A kickout.
- A foul committed by the opposing team.
- A line ball awarded to the team.

It should be noted that the possession events captured in the log data are dependent on the categorical labels defined by the analyst using the Nacsport coding system. Events such as line balls and fouls committed by the team in possession are recorded as turnovers within this coding scheme, meaning they are captured under the turnover category rather than as separate event types. Any possession-changing event not captured by these categories would represent a limitation of the log data and may result in some game phases being incorrectly bounded.

When one of the above events is detected in the log, we extract the click timestamp to mark the start of the game phase. A phase ends when the same team shoots or loses possession, and the corresponding click timestamp is extracted to mark the end of the phase. In the first classification experiment game phases starting when Team A wins the ball from the opponents and ending with a shot are labelled with 1, and all the remaining game phases with 0. In the second experiment, game phases starting when Team B wins the ball from Team A and ending with a shot made by Team B are labelled with 1 and all the remaining game phases with 0. In Figure 6.4 is displayed an example of the extraction and labelling of game phases from the log data.

Category	Start	Click	End
TB Pos from Throw in-up	00:05	00:10	00:15
Throw in	00:05	00:10	00:15
Team A Pos frm TO	00:00	00:15	01:15
Turnover TB	00:15	00:15	00:44
Attack Team A	00:39	00:44	01:05
TA Shot(P)	00:50	00:50	01:05
TB Kickout	01:13	01:13	01:22
TA Pos frm KO	01:12	01:22	01:38
Attack Team A	01:18	01:23	01:38
TA Shot(P)	01:27	01:27	01:38
TB Kickout	01:49	01:49	01:57
TA Pos frm KO	01:47	01:57	02:14
TB Foul	01:55	02:00	02:05
Attack Team A	01:57	02:02	02:14

Start Time	End Time	Description	Exp. 1 Label	Exp. 2 Label
00:15	00:50	TA Shot	1	0
01:22	01:27	TA Shot	1	0
02:02	02:07	TA Loss Ball	0	0
02:07	02:25	TB Shot	0	1

Figure 6.4: Example of extraction and labelling of game phases from log data.

The resulting data frame for the game's phases shows the corresponding events' start and end times. This process is repeated for each of the 11 games in the dataset. A sample of the resulting data frame of the phases of a game is shown in Table 6.1.

Table 6.1: Sample data frame illustrating the phases of the game. Columns include *GameID* (ID of the corresponding game), *Start Event* (event initiating the phase), *End Event* (event concluding the phase), *Start Time* (game time in seconds at the beginning of the phase), *End Time* (game time in seconds at the end of the phase), *Team A Shot* (1 if team A shots 0 otherwise), and *Team B Shot* (1 if team B shots 0 otherwise).

GameID	Start Event	End Event	Start Time	End Time	Team A Shot	Team B Shot
1	Team A Pos frm TO	TA Shot(P)	5	62	1	0
1	TB Kickout	TB Kickout	63	71	0	0
1	TA Pos frm KO	TA Shot(P)	72	98	1	0
1	Team B Pos frm TO	Attack TB	99	149	0	0
1	Team A Pos frm TO	TA Shot(P)	150	227	1	0
1	TB Kickout	TB Shot(P)	228	283	0	1

6.2.2 Feature Extraction from Sensor Data

Following the identification of game phases and their corresponding timestamps, GNSS and IMU sensor data were used to extract features characterizing Team A's collective movement patterns during each phase. Rather than computing aggregate team statistics or individual player metrics, a relational approach was adopted, which quantifies pairwise relationships between all players, represented as adjacency matrices.

As a first step, the raw sensor data from wearable devices were downsampled to 1Hz:

- GNSS data (10Hz): latitude, longitude, and instantaneous speed. Data were downsampled to 1 Hz by averaging every 10 samples within each second.
- IMU data (100Hz): 3D acceleration (a_x, a_y, a_z) and 3D gyroscope (g_x, g_y, g_z). Data were downsampled to 1 Hz by averaging every 100 samples within each second.

For each game phase (defined by start time t_{start} and end time t_{end}), pairwise relationships between every pair of players (i, j) in Team A were computed. For each second t in the interval $[t_{start}, t_{end}]$ delimiting a game phase, the following metrics were calculated:

- *Euclidean distance* $d_{ij}(t)$ between players i and j computed from their latitude and longitude coordinates.

- *Speed difference* $\Delta v_{ij}(t) = |v_i(t) - v_j(t)|$: Absolute difference in instantaneous speed.

- *Acceleration differences* $\Delta a_{ij}^{x}(t)$, $\Delta a_{ij}^{y}(t)$, $\Delta a_{ij}^{z}(t)$: Absolute differences in each acceleration component.

- *Gyroscope differences* $\Delta g_{ij}^{x}(t)$, $\Delta g_{ij}^{y}(t)$, $\Delta g_{ij}^{z}(t)$: Absolute differences in each angular velocity component.

Using absolute values for pairwise differences emphasizes the magnitude of the differences between players, irrespective of direction. For each variable (distance, speed, acceleration x, y, z, gyroscope x, y, z), these sequences of differences across the entire phase duration were aggregated into summary statistics:

1. Average of the pairwise metric over the phase duration.

2. Standard deviation of the pairwise metric over the phase duration.

An example of these steps is shown in Figure 6.5, where second-by-second differences for a single variable are aggregated into summary statistics for one selected game phase and a small sample of three players.

Although only 15 players participated per game phase, metrics were computed for all possible player pairs in Team A across 11 games, since player participation varied between games. Pairwise features involving players not participating in a given phase were assigned a value of -1, to distinguish them from valid values whose minimum is 0 due to absolute difference calculations. This approach ensures the same feature dimensions across all phases.

For each pair of players and game phase, 8 values are generated representing the average differences in distance, speed, acceleration (x, y, z), and gyroscope (x, y, z); and 8 values representing the average standard deviation for distance, speed, acceleration (x, y, z), and gyroscope (x, y, z) during that phase. In other words, for each game phase, 16 37×37 adjacency matrices were generated, where 37 is

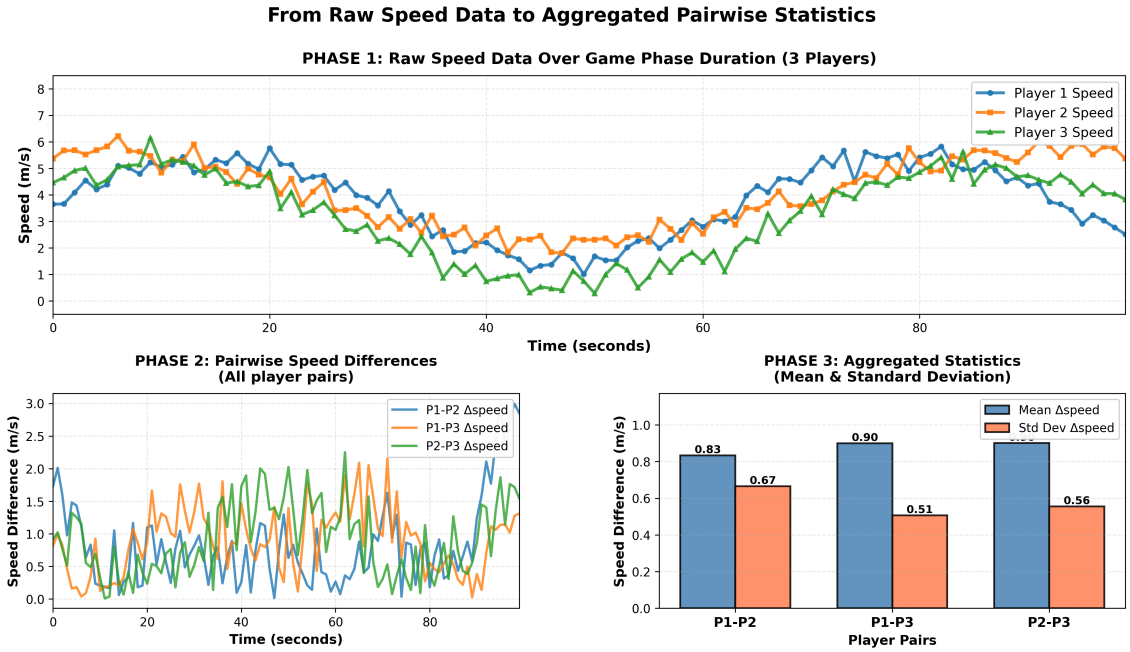


Figure 6.5: Transformation from raw speed data to aggregated pairwise statistics for three players.

the number of players in Team A. Figure 6.6 shows four illustrative matrices for different metrics and a small sub-sample of players.

Then, each adjacency matrix is flattened to realize a long feature vector. In more detail, each column of the final dataset represents the average or standard deviation of the difference of one of the GNSS or IMU variables for a couple of players. For each couple of players, we have 2 (summary statistics) * 8 (sensor variables) = 16 columns. Since we have 37 players in the team, this results in 13,921 columns. Each row represents a phase of game from the 11 games in the sample. This dataset represents the input data for the feature selection and machine learning models.

However, this approach creates a dataset composed of a very high number of columns, with many rows having values equal to -1 in those columns where one of the players in the pair was not playing. This situation is handled in the next steps with a feature selection method.

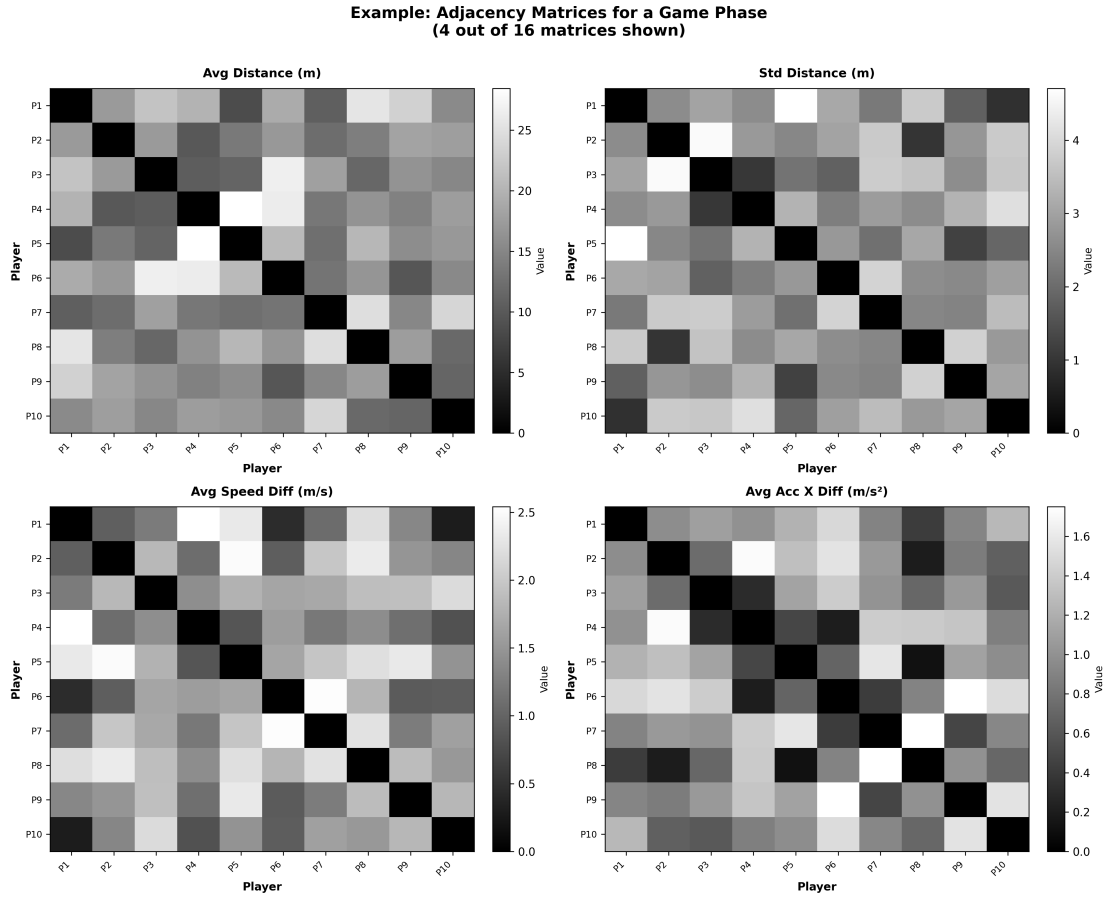


Figure 6.6: Example adjacency matrices representing pairwise metrics between a sample of players during a single game phase. Four of the 16 total matrices are displayed: average Euclidean distance, standard deviation of distance, average speed difference, and average acceleration x difference. Only a small subsample of players is shown.

6.2.3 Model Pipeline

Train-Test Split. The dataset was divided using 10 games for training and 1 game for testing. This approach replaces the leave-one-game-out cross-validation used in the previous chapters, due to computational costs. The automatic feature selection method with cross-validation performed on each model, with 13,921 features, requires several hours. Repeating it across all 11 games would be too expensive. With about 400 training phases from 10 games, the training set is large enough for reliable feature selection, and the single game is used for an objective evaluation.

Recursive Feature Elimination with Cross-Validation. Our approach uses an automated feature selection method to identify the most significant features for predictive modelling from a data frame containing thousands of features. Recursive

Feature Elimination with Cross-Validation (RFECV) method was used, which iteratively removes the least important features and evaluates model performance using k-fold (k = 3) cross-validation [140]. K-fold cross-validation is a method used to assess the performance of an ML model by dividing the data into k equal subsets. The model is trained on k-1 subsets and validated on the remaining subset, and this process is repeated k times, with each subset used exactly once as the validation set. The RFECV method reduces the risk of overfitting and provides a reliable estimate of the model's generalizability. RFECV involves training a model on the full set of features, ranking them based on importance scores, removing the least important feature, and evaluating the model. This cycle repeats until only one feature remains, and the set of features achieving the best cross-validation performance is selected. RFECV was applied only for the training set (10 games). In the testing phase (1 game), each model is evaluated using this optimized feature set to assess its performance.

The motivation for using RFECV in this study is supported by several key factors and references. RFECV is known for its ability to effectively handle high-dimensional datasets by identifying the most relevant features, improving model performance, and interpretability [141]. RFECV method helps in mitigating the risk of overfitting by incorporating cross-validation, which ensures that the selected features generalize well to unseen data [142]. RFECV has been used in various domains, including malicious intrusion detection in computer networks [143], diagnosis of Alzheimer's disease [144], and fraud detection [145]. This demonstrates its robustness in selecting the best feature set.

RFECV was chosen over simpler alternatives such as L1 regularisation for two reasons. L1 regularisation performs feature selection within a linear model by shrinking coefficients to zero, and when features are highly correlated, it tends to arbitrarily select one feature from a correlated group, potentially discarding others that are equally informative [146]. RFECV instead uses the tree-based estimator's own feature importance scores, selecting features consistent with the non-linear decision

boundaries of the final classifier, and directly optimizes the selected subset against cross-validated classification performance [141].

Machine Learning Models. Four machine learning classifiers were evaluated:

- **Logistic Regression**: A probabilistic model that defines a linear decision boundary by fitting a logistic function to the data [147]. For this study, logistic regression is used as a baseline model.
- **Decision Tree**: A non-parametric model that partitions the feature space with binary splits, creating a hierarchical tree structure. At each node, the algorithm selects the feature and split threshold that maximizes information gain [148].
- **Random Forest**: An ensemble method that builds multiple independent decision trees on random subsets of both features and training samples, then aggregates predictions with majority voting [149].
- **XGBoost**: A gradient boosting model that sequentially builds decision trees. Each new tree is trained to correct prediction residuals from all previous trees [102].

Class Imbalance. In both classification tasks, shot phases occur less frequently than all other game phases combined. As a result, the datasets present class imbalance. To address this problem, during model training, undersampling of the majority class was applied [150]. This technique prevents models from achieving high accuracy by simply predicting the majority class [131].

6.3 Results

In this section, the performance of machine learning models are compared on the two distinct classification tasks: identifying game phases that culminate in a shooting event for Team A (Task A) and for the opposing Team B (Task B), using data only from Team A.

This study aims to demonstrate the effectiveness of ML models in classifying game phases, potentially leading to the future automation of event labelling based only on spatio-temporal data from a single team when no log data are available. Both tasks use sensor data collected on the same team during the 11 GF games across the years 2019, 2020, and 2021. Data were collected using the STATSports Apex 10 Hz GNSS and 100Hz IMU device, as described in Section 3.2.1. The analysis includes data from all players participating in these games.

6.3.1 Experiment 1: Team A Shots Classification

This experiment evaluated the ability of machine learning models to classify attacking game phases by Team A that lead to shots, using exclusively wearable sensor data collected from that team.

Logistic Regression, Decision Tree, Random Forest, and XGBoost models were trained and evaluated. For each, RFECV identified the feature subset that achieved the best performance. Table 6.2 summarizes the best results achieved by each model with its optimized features. Random Forest resulted as the top-performing classifier, achieving an F1 score of 0.69 using a substantial feature set (11,136 features). XGBoost showed competitive performance (F1 score 0.67) selecting only three features without compromising accuracy. Logistic Regression's poor performance (F1 score 0.52) suggests the inadequacy of linear models trained using this particular set of features formed by pairwise relationships. Decision Tree used the highest amount of features but underperformed compared to other methods.

The large number of features retained by Random Forest compared to XGBoost reflects a fundamental difference in how the two models handle high-dimensional feature spaces. Random Forest builds each tree independently on a random subset of features and averages predictions across trees, which means it can tolerate a large number of weakly informative features without overfitting, as the averaging process reduces variance. As a result, RFECV finds little cross-validated benefit in removing additional features, since each individual feature contributes marginally across

different trees. XGBoost, by contrast, builds trees sequentially and applies explicit regularisation penalties on the number of leaves and the magnitude of weights [102], which actively discourages the use of redundant features. This makes XGBoost more aggressive in identifying a compact discriminative subset, explaining why it achieved comparable classification performance with only 3 features.

Future work should explore alternative feature selection methods such as LASSO/L1 regularization, which may achieve more aggressive feature pruning than RFECV by penalizing feature coefficients directly and would allow a direct comparison of selected feature subsets and classification performance across methods [146].

Table 6.2: Performance metrics of ML models in classifying Team A shooting events.

Model	**Accuracy**	**F1 Score**	**AUC**	**Precision**	**Recall**	**Selected Features**
Logistic Regression	0.52	0.52	0.53	0.52	0.52	4
Decision Tree	0.52	0.54	0.52	0.52	0.53	13786
Random Forest	**0.63**	**0.69**	0.61	**0.65**	**0.63**	11136
XGBoost	0.61	0.67	**0.67**	0.62	0.61	3

Random Forest was selected for a deeper analysis due to its highest accuracy and F1 score. The confusion matrix for the Random Forest model in Figure 6.7 reveals details about its classification outcomes. The model accurately identified 10 phases that ended in no shot, but it also classified 13 possession intervals incorrectly as shots (false positives). On the other hand, it successfully detected 19 true shooting phases while missing 4 (false negatives). This imbalance, where the model produces more false alarms than missed shots, reflects a modeling choice in this feasibility study to favor recall of shooting phases over precision, given the low base rate of shots relative to non-shot phases.

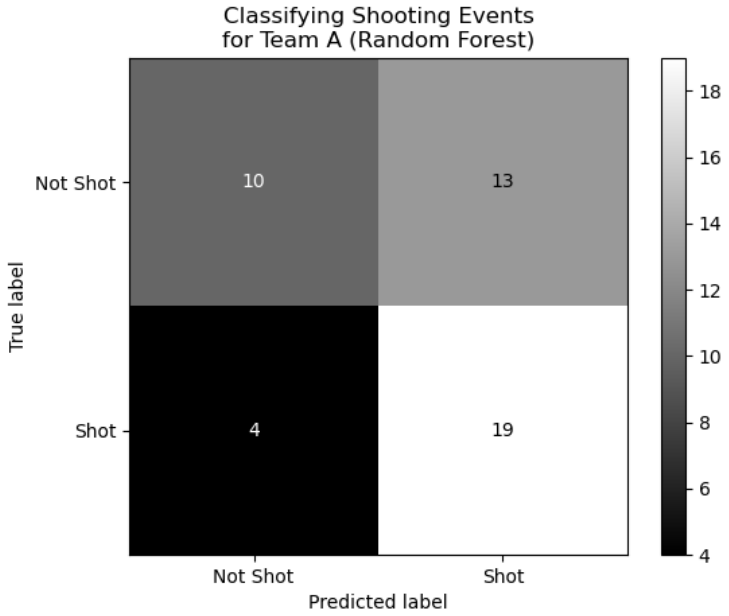


Figure 6.7: Confusion matrix for Random Forest classifying shooting events for Team A.

6.3.2 Experiment 2: Team B Shots Classification

The second experiment aimed to classify game phases ending in shooting events by the opposing team, Team B, using only wearable sensor data from Team A. The methodology is the same as for Experiment 1, involving labelled phases from log data, comprehensive feature engineering aggregating pairwise differences in summary statistics metrics, and recursive feature elimination (RFECV) applied for each model to optimize the set of features.

As shown in Table 6.3, results were similar to experiment 1, with some differences. The Random Forest and XGBoost models resulted as the best models with comparable accuracy scores (0.69). XGBoost achieved the highest AUC (0.78) and precision/recall balance (0.69), while using only 15 features. This finding is similar to Experiment 1, where XGBoost's capacity of using only a small subset of features did not compromise classification performance. Logistic Regression also reached good performance (F1 score 0.68 and AUC 0.70), opposed to the result in Experiment 1. In contrast, Decision Tree performance was notably weaker across metrics, underscoring the benefits of ensemble approaches in handling complex and

high-dimensional data typical of wearable sensor data.

The confusion matrix for the XGBoost model, illustrated in Figure 6.8, shows 19 true negatives, 7 false positives, 9 false negatives, and 16 true positives. Compared to Experiment 1, this confusion matrix exhibits a more balanced false positive to false negative ratio, indicating the model struggles more in identifying all shooting phases but effectively reduces false alarms.

Table 6.3: Performance metrics of ML models in predicting opponent (Team B) shooting events.

Model	Accuracy	F1 Score	AUC	Precision	Recall	Selected Features
Logistic Regression	0.67	0.67	0.70	0.67	0.67	1695
Decision Tree	0.47	0.45	0.47	0.47	0.47	8294
Random Forest	**0.69**	**0.68**	0.73	0.68	0.68	11325
XGBoost	**0.69**	**0.68**	**0.78**	**0.69**	**0.69**	15

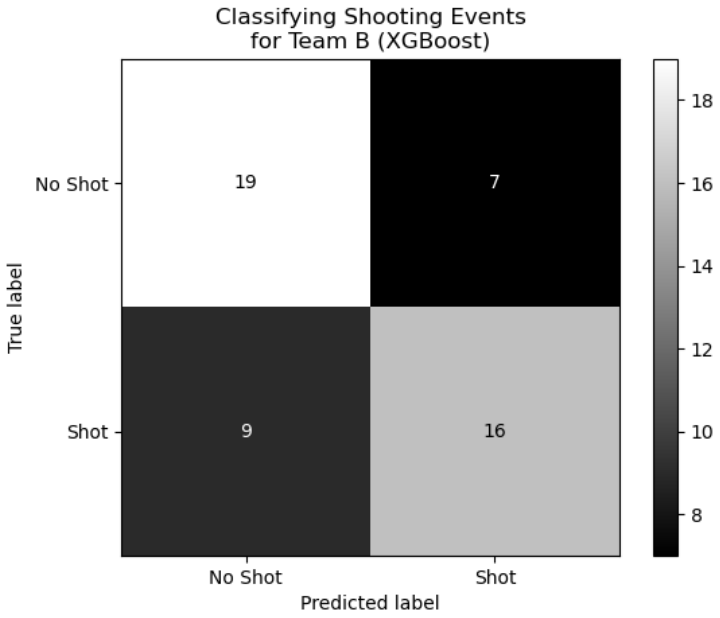


Figure 6.8: Confusion matrix for XGBoost classifying shooting events for Team B.

6.4 Summary

In this chapter, the potential for automatic classification of game phases using sensor data from only one team was explored. In particular, two experiments were conducted: one aimed to identify game phases ending with shots made by the team

with available data, and the other focused on identifying shots made by the opposing team.

A new feature engineering approach was defined for both classification tasks. Each feature in the input dataset for the classifiers represents the average or standard deviation of differences in several variables (distance, speed, 3D acceleration, 3D gyroscope) between each pair of players in the team during each game phase. This approach constitutes a new method to construct features for machine learning models in sports, which looks at pairwise differences in variables between players represented as adjacency matrices.

Several machine learning classifiers were tested for both experiments. Due to the large number of features, a feature selection process was implemented, applying recursive feature elimination with cross-validation for each model to identify the feature set that maximized the F1 score.

All tested models achieved good classification performance in both experiments. For classifying Team A shot phases, Random Forest achieved the highest performance (accuracy 0.63, F1 score 0.69) using a large feature set (11,136 features), while XGBoost obtained similar results (accuracy 0.61, F1 score 0.67) with fewer features (3). A similar situation was observed for Team B shot phases: both XGBoost and Random Forest produced comparable results (accuracy 0.69, F1 score 0.68) but with substantial differences in features' use (15 vs. 11,325).

These results establish a proof-of-concept baseline demonstrating that collective inter-player movement patterns from a single team's wearable sensor data carry discriminative information about game phases ending in shots, without requiring ball position, opponent tracking, or manual video annotation. However, the findings should be interpreted with caution. The models were evaluated on data from a single team across 11 games, which limits generalisability, and the F1 scores indicate that a proportion of phases are still misclassified. Further validation across multiple teams, competitions, and seasons is needed before these models could be considered reliable for applied use. The log data used for labelling were produced by a single analyst,

introducing the possibility of labelling inconsistency, and the high dimensionality of the feature space relative to the available sample size may affect model stability despite the feature selection procedure. From a practical standpoint, the value of this framework lies in its potential to reduce the manual annotation burden currently placed on video analysts. A match currently requires hours of post-match coding to generate the contextual log data that clubs use alongside sensor data. By pre-labelling candidate shot phases automatically, this framework could reduce that workload substantially, and for clubs without dedicated analyst staff, it could make basic tactical context extraction feasible where it is currently not possible. These benefits are prospective and contingent on further validation. The framework is also not limited to Gaelic football, since the approach relies on pairwise inter-player movement relationships, it is transferable to any other invasion-field team sport where wearable sensor data are collected.

Chapter 7

Conclusion

7.1 Dissertation Overview

This thesis addressed four fundamental challenges in sports analytics: the absence of transparent methodologies for processing sensor data, limited understanding of spatio-temporal movement patterns, lack of studies focusing on predicting future metabolic workload, and dependence on expensive manual video analysis for extracting tactical context. Four distinct frameworks were developed to address these challenges using wearable sensor data from 11 elite Gaelic Football matches (76,313,030 raw observations from STATSports Apex 10 Hz GNSS with 100 Hz IMU).

Chapter 3 presented a systematic feature engineering framework that transforms raw sensor data into structured Action and Event datasets through documented algorithms, achieving 99.8% dimensionality reduction while preserving analytical information. Chapter 4 introduced the Time-Window Spatial Activity Graph framework, using spatial cells as graph nodes to reveal critical transition zones and spatial communities with distinct workload profiles. Chapter 5 developed a predictive framework for forecasting 3-minute metabolic workload transitions, introducing the Metabolic Work Ratio metric and achieving accuracy 0.71 ± 0.04, and F1 scores of 0.69 ± 0.05. Chapter 6 demonstrated that shooting phases of both teams can be automatically classified with 0.63-0.69 accuracy using only single-team sensor data through adjacency matrices of pairwise player movement relationships as machine

learning input features.

These frameworks represent parallel analytical approaches rather than a sequential pipeline. The Action dataset from Chapter 3 is used as input directly only into Chapter 4 for spatial analysis, while Chapters 5 and 6 extract features directly from raw data at different temporal granularities suited to their specific objectives.

7.2 Research Questions and Findings

This section summarizes the key findings for each research question, demonstrating how the four frameworks address the identified challenges in sports analytics.

7.2.1 Chapter 3: Developing a Threshold-Based Activity Feature Set

The first research question asked whether raw sensor data could be systematically transformed into structured datasets that preserve movement information while enabling diverse analyses. The developed framework uses velocity zone thresholds from Rampinini et al. [53] to segment continuous data into discrete actions (continuous movement within one velocity zone) and aggregates these into events (sequences separated by rest periods). The framework reduced 76 million raw observations into 148,371 actions and 54,821 events across 11 matches.

Validation confirmed information preservation: derived running profiles matched literature benchmarks with players covering 8,721 ± 1,299 m total distance (vs. 8,160 ± 1,482 m in [23]), 1,596 ± 543 m high-speed distance (vs. 1,731 ± 659 m), and 462 ± 178 m sprint distance (vs. 445 ± 169 m). Actions averaged 2.9 seconds in duration with 2.55 actions per event, and events averaged 7.1 seconds in duration.

The practical utility was demonstrated through predictive modeling of second-half high-speed running distance. XGBoost models achieved an average error of 65 meters per player per half. The framework successfully addresses the transparency gap by providing complete algorithmic specifications with pseudocode, published in

[49], [50].

7.2.2 Chapter 4: Interpreting Player Movement Using a Spatial Grid Model

The second research question examined whether graph representations with spatial cells as nodes could reveal movement patterns that traditional player-centric analyses cannot capture. Existing approaches in team sports use graph theory to model player-to-player interactions through passing networks, where players are nodes. The Time-Window Spatial Activity Graph framework introduces a fundamentally different approach: spatial cells become nodes, with edges representing player transitions between locations. This enables identification of critical spatial zones, detection of areas with similar movement characteristics, and analysis of how spatial usage evolves during matches. These insights cannot be obtained from player networks or traditional position tracking.

The framework divides playing areas into cells through tessellation algorithms, maps player coordinates to the nearest cells, and constructs directed multi-graphs where edge metadata captures transition frequency, speed, acceleration, and intensity distributions. Two case studies demonstrated the framework's capabilities using different temporal approaches: Game 811 used rolling 5-minute windows to track continuous evolution, while Game 893 used fixed 15-minute windows to compare distinct match phases.

Betweenness centrality analysis in Game 811 revealed how critical zones shift during matches. In the opening 5 minutes, highest centrality scores concentrated in the central corridor where players connected defensive and attacking moves. By minutes 3-8, centrality shifted toward the attacking region near the opposition goal. This demonstrates that the framework identifies cells acting as bridges in movement sequences, not just areas with high activity volume.

Community detection using the Louvain algorithm revealed spatial clusters where players moved together with distinct characteristics. In Game 893's first 15 minutes,

nine communities were identified. Community 1 (offensive and central-right, 37 cells, 14 players) showed 2.93 m/s average speed with 82 high-acceleration events and 6% high-intensity actions. Community 7 (defensive-left, 18 cells, 14 players) exhibited the highest speed at 3.02 m/s with 8% high-intensity actions, indicating a corridor for fast defensive or counter-attacking movements. In contrast, Community 2 (central and defensive center-right, 28 cells, 14 players) displayed 2.74 m/s average speed with 5% high-intensity actions, reflecting more measured pacing.

Temporal analysis between the two 15-minute windows revealed structural simplification from nine to six communities, indicating reduced distances between player lines. The midfield-offensive community (Community 0 in the second window: 39 cells, 14 players) maintained elevated activity with 3.0 m/s average speed, 115 high-acceleration events, and 6.7% high-intensity actions. The defensive-left community (Community 1: 37 cells, 14 players) showed 2.75 m/s with 82 acceleration events and 6% high-intensity actions. These patterns persisted across windows: midfield and offensive regions consistently demanded higher speeds and more frequent accelerations, while the left side maintained high-speed activity presence and the right side exhibited less pronounced intensity.

The framework's contribution, presented in [51], is methodological: it provides the first cell-based spatial graph approach in team sports that integrates location, movement intensity, and temporal evolution into a unified analytical model. This enables quantification of spatio-movement patterns and identification of critical transition zones.

7.2.3 Chapter 5: Forecasting Transitions in Metabolic Workload

The third research question investigated whether machine learning could predict imminent workload transitions to enable real-time interventions. The framework operates on 3-minute rolling windows, introducing the Metabolic Work Ratio (ratio of high metabolic load distance to total distance, with high load defined as metabolic

power $\geq$ 20 W/kg). The prediction task classifies whether players will experience significant MWR changes (exceeding 75th percentile) in the next 3-minute window.

Feature engineering extracted different features per window: features describing the activity in the current window (velocity zones, metabolic power metrics, accelerations, spatial distances to teammates) and cumulative features summarizing activity from the start of the match up to the end of the current window (for example, total distance covered and total high-speed distance). Gradient Boosting achieved the best performance with accuracy 0.71 $\pm$ 0.04, and F1 score 0.69 $\pm$ 0.05 across all players and matches, substantially exceeding baseline performance.

Performance varied with player participation: high-presence players (174, 146, 201) showed stable F1 scores of 0.65-0.75, while low-presence players exhibited lower scores (0.45-0.60), indicating models learn player-specific patterns with sufficient data. Temporal analysis revealed F1 scores generally increased toward the end of each half, suggesting prediction improves as historical data accumulates.

SHAP analysis identified key predictive factors: cumulative metrics (metabolic power, distance, and high metabolic load distance) were strongest positive predictors, while percent time in metabolic power zone 4 and maximum speed in the current window showed strong negative associations. This pattern captures work-recovery dynamics where high-intensity efforts are followed by recovery phases. The framework demonstrates that 3-minute forward prediction achieves meaningful accuracy for proactive interventions when real-time data streams are available.

7.2.4 Chapter 6: Detecting Shot Events in Game Phases

The fourth research question investigated whether shooting events could be automatically classified using only single-team sensor data, eliminating dependence on ball tracking or manual video analysis. The framework focuses on game phases (possession sequences from ball acquisition to loss or shot) ending in shooting events due to their tactical significance and distinctive movement signatures.

The novel feature engineering approach computes pairwise player differences: for

each phase, summary statistics (mean, standard deviation) of six variables (distance, speed, 3D accelerometer, and 3D gyroscope) are calculated per player. Then these differences between every player pair are computed and represented as adjacency matrices, which serve as input features for machine learning classifiers.

This generates 13,950 features per phase, encoding inter-player coordination patterns. Recursive feature elimination with cross-validation identifies optimal feature subsets for each model.

Two experiments were conducted. For own-team shooting phases (Experiment 1), Random Forest achieved 0.63 accuracy and 0.69 F1 score using 11,136 features, correctly identifying 19 of 23 actual shots (83% recall) with the confusion matrix showing 10 true negatives, 13 false positives, 19 true positives, and 4 false negatives. The bias toward false positives reflects the model prioritizing shot detection over precision. XGBoost achieved comparable performance (0.61 accuracy, 0.67 F1) using only 3 features.

For opponent shooting phases (Experiment 2), both XGBoost and Random Forest achieved 0.69 accuracy and 0.68 F1 score. XGBoost used just 15 features, with the confusion matrix showing 19 true negatives, 7 false positives, 16 true positives, and 9 false negatives. The ability to detect opponent shots using only own-team defensive movements demonstrates that coordination patterns contain sufficient signal about opponent tactical situations. This work, presented in [52], establishes the feasibility of automatic phase labeling that could reduce manual video analysis requirements.

7.3 Integrated Findings and Practical Implications

The four frameworks developed in this thesis address distinct but related analytical challenges, and understanding how they relate to each other analytically is important for interpreting the contribution of the thesis as a whole. Chapter 3 takes raw sensor data as input and produces two structured datasets (the Action dataset and the Event dataset) through a reproducible, documented pipeline. The Action dataset is

then used directly as input into Chapter 4 for spatial graph construction. Chapters 5 and 6 both work from the original raw sensor data, extracting features at different temporal granularities suited to their specific prediction objectives. The frameworks therefore do not form a strict sequential pipeline, but they share a common data foundation and address complementary dimensions of the same underlying problem: how to extract more value from wearable sensor data already collected by elite Gaelic football clubs.

After conducting this programme of research, what is now known is that wearable GNSS and IMU sensor data, when processed through transparent, reproducible frameworks, can support a substantially broader range of analytical objectives than the current sports science literature exploits. Physical movement data can be structured into interpretable datasets that preserve temporal and intensity information while reducing computational burden by 99.8%. Spatial representations of movement reveal how different areas of the playing field are used at different intensities across match phases, providing a perspective on performance that aggregate running profiles cannot offer. Machine learning models trained on engineered features from these data can forecast short-term metabolic workload transitions with meaningful accuracy, offering a basis for proactive in-match decision support. Finally, collective inter-player movement patterns carry sufficient signal to classify game phases ending in shots (for both the tracked team and their opponents) without ball position data or manual annotation. Collectively, these findings suggest that a single team's sensor data contains richer information than is currently extracted in applied practice.

7.4 Thesis Limitations

Several limitations apply across the thesis as a whole and should be considered when interpreting the findings.

Single team and limited sample. All four frameworks were developed and evaluated using data from a single elite inter-county Gaelic football team across 11 matches. While this provides a consistent and ecologically valid dataset, it limits

the generalizability of all findings. Spatial patterns, workload transition dynamics, and shot phase classification signals may reflect team-specific tactical organization rather than universal characteristics of the sport. Replication across multiple teams, competition levels, and seasons is necessary before any framework can be considered validated for broad application.

Single season and contextual factors. The 11 matches span three years but represent a single team's competitive record. Match-level contextual factors (scoreline, opposition quality, tactical formation, weather, and pitch dimensions) vary across games and are not controlled for in any analysis. These factors may substantially influence the physical and tactical patterns observed, and their absence from the models represents an unobserved source of variability.

GPS and IMU measurement limitations. All analyses depend on GNSS and IMU data collected at 10 Hz and 100 Hz respectively using the STATSports Apex device. GPS accuracy is reduced at high velocities and during non-linear movements, which introduces measurement uncertainty into velocity zone classifications, metabolic power estimates, and pairwise distance calculations.

Temporal cross-validation constraints. The Leave-1-Game-Out cross-validation scheme used in Chapters 5 and 6 does not guarantee strict temporal ordering across all folds. Only the fold where the chronologically last game is held out preserves full temporal separation. The small dataset size makes a fully chronological holdout impractical, and results should be interpreted as indicative of model potential rather than definitive performance estimates.

Log data limitations. The tactical context data used in Chapter 6 were generated by a single trained analyst using Nacsport software. Inter-rater reliability was not assessed, and inconsistencies in event coding may have introduced noise into the classification labels.

Proof-of-concept status. All four frameworks are presented as proof-of-concept contributions rather than validated applied tools. The MWR metric's physiological validity has not been established through correlation with objective measures

such as lactate or VO_2. The spatial patterns identified in Chapter 4 are case-study observations from two games. The shot classification models in Chapter 6 achieve promising but imperfect F1 scores on a small test set. Each framework requires substantially larger and more diverse datasets before it can be recommended for routine applied use.

7.5 Suggestions for Future Research

This final section of the dissertation examines opportunities for future work based on this research:

1. **Advanced Graph Analysis:** Extend spatial graph framework through graph neural networks (GNNs) and predictive modeling to forecast spatial organization. The cell-based graph representation from Chapter 4 could serve as input to temporal GNNs trained to predict future spatial configurations: given spatial graphs from minutes 0-15, predict the likely high-traffic zones and community structures for minutes 15-30. Integrating event log data (shot attempts, turnovers, scores) would enable tactical outcome prediction. For example, forecasting whether current spatial patterns precede goal-scoring opportunities or poor defensive scenarios within the next 3-5 minutes. This would transform spatial analysis from post-match description into predictive decision support, enabling coaches to anticipate tactical developments and suggest proactive adjustments.

2. **MWR Validation:** Correlate Metabolic Work Ratio with VO_2 measurements during controlled training, examine MWR-lactate relationships validating intensity classifications, assess whether MWR changes predict autonomic stress through heart rate variability, and conduct prospective studies examining whether MWR patterns precede injury events. While the MWR metric introduced in Chapter 5 is theoretically grounded in metabolic power frameworks and demonstrates predictive utility (F1 score 0.69 for transition fore-

casting), its physiological validity remains unestablished.

3. **Automatic Tactical Role Identification and Player Development:** Use unsupervised learning on movement patterns and spatial usage to automatically discover playing styles and track tactical evolution. Beyond simply clustering players into traditional positions, this approach could identify emerging new tactical roles (for example, "inverted wing-back," "box-to-box midfielder," "pressing forward") based purely on observable movement signatures—spatial heat maps, velocity distributions, and transition patterns extracted from the Action dataset (Chapter 3). Tracking individual players across multiple matches would reveal tactical flexibility: does a player consistently occupy one role, or do they switch between multiple tactical profiles depending on opponent or match situation? This may result in direct scouting applications (identifying players with rare tactical profiles), player management implications, and player development value (tracking whether young players' movement patterns resemble more experienced elite athletes).

4. **Real-Time Streaming Data Architecture:** Develop technical infrastructure for live predictive analytics during matches. Current STATSports systems provide data only after matches conclude, but the predictive models in Chapters 5 and 6 demonstrate potential for real-time decision support. Implementation requires streaming sensor data wirelessly to sideline computers that immediately compute features (velocity zones, metabolic power, spatial positions) and run trained models to generate predictions within seconds. The models must be computationally efficient: the Gradient Boosting models from Chapter 5 can generate predictions in milliseconds on standard laptops, making real-time deployment technically feasible. The system would need fault tolerance: if a sensor temporarily disconnects, the system should continue operating with available data rather than failing. A practical deployment might provide coaches with tablet displays showing which players are predicted to enter high metabolic load in the next minutes and whether current movement

patterns suggest imminent opponent shooting opportunity.

The frameworks established in this thesis provide methodological foundations upon which these research directions can systematically build, advancing sports analytics from descriptive pattern identification toward predictive and prescriptive performance optimization.

Appendix A

Terminology

This section defines key terms used throughout the thesis to ensure clarity and consistency.

Raw Sensor Data refers to the unprocessed time-series measurements directly recorded by wearable sensors, including GNSS position and velocity data sampled at 10 Hz and IMU accelerometer and gyroscope data sampled at 100 Hz. Each observation represents a single measurement without aggregation or transformation, returned by the sensor software.

Structured Dataset refers to a tabular data representation where each observation corresponds to a movement segment defined by specific criteria (for example, a continuous movement within a velocity zone) rather than individual sensor readings. In this context, structured datasets are characterized by dimensional reduction from high-frequency sensor data into discrete units and summary statistics and features computed for each unit.

Features or **Metrics**. In this research, the terms *features* and *metrics* are used interchangeably to refer to computed attributes derived from raw sensor data through mathematical operations. "Features" often refer to columns in datasets or input variables for machine learning models and "metrics" refer to performance measurements for analysis, both describe the same underlying computed attributes in this thesis.

Framework refers to a structured methodological approach that defines the

processes, algorithms, and data transformations required to accomplish a specific analytical objective. In this thesis, frameworks encompass the complete workflow from raw data acquisition through transformation to final analysis, including the Action-Event framework (Chapter 3) and the Spatial Grid framework (Chapter 4).

Pipeline refers to a sequence of data processing steps where the output of one step becomes the input for the next, creating an automated workflow for analysis or prediction. Pipelines in this thesis include feature engineering pipelines that transform raw sensor data into analytical features and machine learning pipelines that combine preprocessing, feature selection, model training, and prediction steps into reproducible workflows.

Data Science Techniques refers to computational methods that combine statistical analysis, machine learning, and domain knowledge to extract insights from data. In the context of this thesis, data science techniques encompass feature engineering, supervised and unsupervised machine learning, time series analysis, graph-based analysis, and predictive modelling applied to sports sensor data.

Data Mining refers to the process of discovering patterns, correlations, and meaningful structures in large datasets through algorithmic methods. In this research, data mining techniques are applied to identify movement patterns, detect spatial communities, extract characteristic sequences, and uncover relationships between player movements and match contexts.

Velocity Zones are discrete speed categories used to classify movement intensity.

Intensity refers to the level of physical effort or exertion during movement, typically quantified through velocity zones, acceleration magnitudes, or metabolic power metrics.

Action is defined formally in Chapter 3 as a temporal interval $[t_s, t_e]$ during which a player's velocity remains within the same velocity zone. Each action is characterized by features including duration, distance covered, maximum speed, and acceleration statistics.

Event is defined formally in Chapter 3 as a sequence of consecutive actions separated from other events by periods where velocity remains below 2 m/s for a minimum duration of one second. Events represent higher-level movement patterns and are characterized by aggregate features computed across their constituent actions.

Movement Unit is a general term encompassing any discrete segment of player movement extracted from continuous sensor data. Actions and events are specific types of movement units with distinct segmentation criteria.

Segmentation refers to the algorithmic process of partitioning continuous sensor data into discrete movement units based on velocity zone transitions. The segmentation algorithm identifies temporal boundaries where a player's velocity transitions between zones, creating action-level observations.

Aggregation refers to the process of grouping consecutive movement segments into higher-level sequences based on temporal proximity and intensity thresholds. The aggregation algorithm combines actions into events by identifying separating periods of rest or low-intensity movement.

Rest or **Recovery Period** refers to intervals where player velocity remains below 2 m/s, indicating periods of low physical demand between higher-intensity movements.

Workload and **Running Performance**. While these terms emphasize different aspects, they are used to describe the same underlying measurements in this research. Both define the cumulative physical demands placed on a player. They are quantified through metrics such as total distance, high-speed distance, sprint counts, accelerations, and duration of activity.

Spatio-Temporal Movement refers to player movement patterns analyzed simultaneously across both spatial dimensions (location on the playing field) and temporal dimensions (changes over time). Spatio-temporal analysis captures where players move, when movements occur, and how movement patterns evolve throughout a match, enabling identification of tactical patterns, territorial control, and

dynamic team behaviors.

Transition Zones refers to spatial regions or cells on the playing area where players frequently change velocity zones, direction, or movement intensity. In graph-based spatial analysis (Chapter 4), transition zones are identified as high-betweenness nodes that serve as critical pathways connecting different areas of the field.

Time Window refers to a fixed duration interval used for aggregating or analyzing sensor data. Time windows are commonly used to compute rolling statistics or segment continuous data into temporal units for analysis.

Performance in the context of sports analytics refers to the quantifiable aspects of player output and capability during training or competition. Performance can be measured through multiple dimensions, including physical output (distance, speed, and acceleration) and technical execution (skills and tactics). In this thesis, performance primarily focuses on physical and movement-based metrics derived from sensor data.

Accuracy measures the proportion of correctly classified instances among all instances. It is defined as:

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

where TP is true positives, TN true negatives, FP false positives, and FN false negatives.

Precision is the ratio of correctly predicted positive observations to the total predicted positives:

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

Recall is the ratio of correctly predicted positive observations to all actual positives:

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

F1 Score is the harmonic mean of precision and recall, providing a balance

between the two metrics:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix is a tabular summary showing the counts of true positive, true negative, false positive, and false negative predictions, used to visualize the performance of a classification model.

AUC (Area Under the Curve) is the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) over different threshold values. It is approximated by:

$$\text{AUC} \approx \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \times \frac{TPR_i + TPR_{i+1}}{2}$$

where TPR_i and FPR_i are the true positive and false positive rates at the i^{th} threshold. This metric indicates the model's ability to discriminate between positive and negative classes, with values closer to 1 representing better discrimination.

Bibliography

[1] Grand View Research. *Sports Analytics Market Size, Share & Trends Analysis Report*. https://www.grandviewresearch.com/industry-analysis/sports-analytics-market. Market size estimated at USD 4.47 billion in 2024, projected to reach USD 14.48 billion by 2030. 2024.

[2] Zone7 Technologies. *Zone7 Expands Service to Liverpool FC*. https://zone7.ai/news/client-announcements/zone7-expands-service-to-liverpool-fc/. Liverpool FC reported 33% reduction in days lost to injury after implementing Zone7's AI-based system. 2022.

[3] Leaders Performance Institute. *How AI Fits into Liverpool's Multidisciplinary Monitoring Model*. https://leadersinsport.com/performance-institute/articles/how-ai-fits-into-liverpools-multidisciplinary-monitoring-model/. Interview with Conall Murtagh discussing 33% reduction in days lost and 40% reduction in substantial injuries. 2022.

[4] Elia Morgulev, Ofer H Azar, and Ronnie Lidor. "Sports analytics and the big-data era". In: *International Journal of Data Science and Analytics* 5.4 (2018), pp. 213–222.

[5] A Mansurali and Ali B Mahmoud. "Sports analytics: Data-driven sports and decision intelligence". In: *Sports Analytics: Data-Driven Sports and Decision Intelligence*. Springer, 2024, pp. 1–17.

[6] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[7] Branch Rickey. “Goodby to some old baseball ideas”. In: *Life* 2 (1954), pp. 78–89.

[8] Bill James. *The new Bill James historical baseball abstract*. Simon and Schuster, 2010.

[9] Ray C Fair and Danielle Catambay. “Branch Rickey’s equation fifty years later”. In: *NINE: A Journal of Baseball History and Culture* 17.1 (2008), pp. 111–118.

[10] Mike D Hughes and Roger M Bartlett. “The use of performance indicators in performance analysis”. In: *Journal of sports sciences* 20.10 (2002), pp. 739–754.

[11] James Mead, Anthony O’Hare, and Paul McMenemy. “Expected goals in football: Improving model performance and demonstrating value”. In: *Plos one* 18.4 (2023), e0282295.

[12] Radivoj Mandić, Saša Jakovljević, Frane Erčulj, and Erik Štrumbelj. “Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017”. In: *PloS one* 14.10 (2019), e0223524.

[13] Macfarlane TU Scott, Tannath J Scott, and Vincent G Kelly. “The validity and reliability of global positioning systems in team sport: a brief review”. In: *The Journal of Strength & Conditioning Research* 30.5 (2016), pp. 1470–1490.

[14] Auksė Miškinytė and Audrius Dėdelė. “Objective assessment of physical activity patterns based on accelerometer and GPS data in adults”. In: *Travel Behaviour and Society* 25 (2021), pp. 112–119.

[15] Thomas Bennett, Phil Marshall, Steve Barrett, James J Malone, and Christopher Towlson. “Quantifying high-speed running in rugby league: An insight into practitioner applications and perceptions”. In: *International Journal of Sports Science & Coaching* (2022), p. 17479541221112825.

[16] Dan B Dwyer and Tim J Gabbett. "Global positioning system data analysis: Velocity ranges and a new definition of sprinting for field sport athletes". In: *The Journal of Strength & Conditioning Research* 26.3 (2012), pp. 818–824.

[17] Alessio Rossi, Enrico Perri, Athos Trecroci, Marco Savino, Giampietro Alberti, and F Marcello Iaia. "GPS data reflect players' internal load in soccer". In: *2017 IEEE international conference on data mining workshops (ICDMW)*. IEEE. 2017, pp. 890–893.

[18] M Mandorino, AJ Figueiredo, G Cima, and A Tessitore. "Analysis of relationship between training load and recovery status in adult soccer players: a machine learning approach". In: *Journal homepage: http://iacss. org/index. php? id* 21.2 (2022).

[19] Shane Malone, Barry Solan, and Kieran Collins. "The running performance profile of elite Gaelic football match-play". In: *The Journal of Strength & Conditioning Research* 31.1 (2017), pp. 30–36.

[20] David M Corbett, Alice J Sweeting, and Sam Robertson. "A change point approach to analysing the match activity profiles of team-sport athletes". In: *Journal of sports sciences* 37.14 (2019), pp. 1600–1608.

[21] Lorena Torres-Ronda, Emma Beanland, Sarah Whitehead, Alice Sweeting, and Jo Clubb. "Tracking systems in team sports: a narrative review of applications of the data and sport specific analysis". In: *Sports Medicine-Open* 8.1 (2022), p. 15.

[22] Michael Mondello and Christopher Kamke. "The introduction and application of sports analytics in professional sport organizations". In: *Journal of Applied Sport Management* 6.2 (2014), p. 11.

[23] Shane Malone, Barry Solan, Kieran D Collins, and Dominic A Doran. "Positional match running performance in elite Gaelic football". In: *The Journal of Strength & Conditioning Research* 30.8 (2016), pp. 2292–2298.

[24] Martin Ryan, Shane Malone, and Kieran Collins. "Acceleration profile of elite Gaelic football match play". In: *The Journal of Strength & Conditioning Research* 32.3 (2018), pp. 812–820.

[25] Thomas Dekkers, Kieran O'Sullivan, Catherine Blake, Joseph G McVeigh, and Kieran Collins. "Epidemiology and moderators of injury in Gaelic football: a systematic review and meta-analysis". In: *Journal of science and medicine in sport* 25.3 (2022), pp. 222–229.

[26] Thomas Reilly and Kieran Collins. "Science and the Gaelic sports: Gaelic football and hurling". In: *European Journal of Sport Science* 8.5 (2008), pp. 231–240.

[27] T Reilly and D Doran. "Science and Gaelic football: A review". In: *Journal of Sports Sciences* 19.3 (2001), pp. 181–193.

[28] Martin Ryan, Shane Malone, Andrew Donnellan, and Kieran Collins. "Acceleration profile of elite Gaelic football with special reference to position of play". In: *The Journal of Strength & Conditioning Research* 34.6 (2020), pp. 1750–1758.

[29] Brian Cunniffe, Wayne Proctor, Julien S Baker, and Bruce Davies. "An evaluation of the physiological demands of elite rugby union using global positioning system tracking software". In: *The Journal of Strength & Conditioning Research* 23.4 (2009), pp. 1195–1203.

[30] Garrett F Coughlan, Brian S Green, Paul T Pook, Eoin Toolan, and Sean P O'Connor. "Physical game demands in elite rugby union: a global positioning system analysis and possible implications for rehabilitation". In: *Journal of orthopaedic & sports physical therapy* 41.8 (2011), pp. 600–605.

[31] Ben Wisbey, Paul G Montgomery, David B Pyne, and Ben Rattray. "Quantifying movement demands of AFL football using GPS tracking". In: *Journal of science and Medicine in Sport* 13.5 (2010), pp. 531–536.

[32] Samuel J Janetzki, Pitre C Bourdon, Kevin I Norton, Jackson C Lane, and Clint R Bellenger. "Evolution of physical demands of Australian football league matches from 2005 to 2017: a systematic review and meta-regression". In: *Sports Medicine-Open* 7.1 (2021), p. 28.

[33] Kevin McGuigan, Mike Hughes, and Denise Martin. "Performance indicators in club level Gaelic football". In: *International Journal of Performance Analysis in Sport* 18.5 (2018), pp. 780–795.

[34] Lorcan S Daly, Ciarán Ó Catháin, and David T Kelly. "Gaelic football match-play: performance attenuation and timeline of recovery". In: *Sports* 8.12 (2020), p. 166.

[35] Shane Mangan and Kieran Collins. "A rating system for Gaelic football teams: Factors that influence success". In: *Journal homepage: http://iacss. org/index. php? id* 15.2 (2016).

[36] Arpad E Elo and Sam Sloan. "The rating of chessplayers: Past and present". In: *(No Title)* (1978).

[37] Aidan J Brady, Michael Scriney, Niall M Moyna, and Andrew McCarren. "Identification of movement categories and associated velocity thresholds for elite Gaelic football and hurling referees". In: *International Journal of Performance Analysis in Sport* 21.5 (2021), pp. 741–753.

[38] Kevin McGuigan, Kieran Collins, and Kevin McDaid. "How do situational variables, related to shooting, in elite Gaelic football impact shot success?" In: *International Journal of Sports Science & Coaching* 17.5 (2022), pp. 1128–1136.

[39] Kevin McDaid, Kevin McGuigan, Jack McDonnell, and Kieran Collins. "An expected shot outcome model for points in elite Gaelic football". In: *International Journal of Performance Analysis in Sport* 25.2 (2025), pp. 175–189.

[40] Dermot Sheridan, Aidan J Brady, Dongyun Nie, and Mark Roantree. "Predictive analysis of ratings of perceived exertion in elite Gaelic football". In: *Biology of Sport* 41.4 (2024), pp. 61–68.

[41] Mauricio C Cordeiro, Ciaran O Cathain, Lorcan Daly, David T Kelly, and Thiago B Rodrigues. "A synthetic data-driven machine learning approach for athlete performance attenuation prediction". In: *Frontiers in Sports and Active Living* 7 (2025), p. 1607600.

[42] Mauricio C Cordeiro, Ciaran O Cathain, Lorcan Daly, David T Kelly, and Thiago B Rodrigues. "A Machine Learning-Based Performance Monitoring Approach for Athlete Performance Attenuation Prediction". In: (2025).

[43] Kate Timmons, Kieran Collins, and Shane Mangan. "The science of Gaelic football match-play and recommendations for future research directions: a narrative review". In: *Sport Sciences for Health* 21.4 (2025), pp. 2505–2521.

[44] Eurisy. *Tracking GAA Players with GPS*. `https://www.eurisy.eu/news/tracking-gaa-players-with-gps_218`. 2021.

[45] STATSports. *LGFA: Leading the Way in Player Performance and Welfare*. `https://statsports.com/news/lgfa-leading-the-way-in-player-performance-and-welfare/`. 2024.

[46] Cloe Cummins, Rhonda Orr, Helen O'Connor, and Cameron West. "Global positioning systems (GPS) and microtechnology sensors in team sports: a systematic review". In: *Sports medicine* 43 (2013), pp. 1025–1042.

[47] Shane Malone, Aidan Shovlin, Kieran Collins, Allistar McRobert, and Dominic Doran. "Is the metabolic power paradigm ecologically valid within elite Gaelic football?" In: *Sport Sciences for Health* 17 (2021), pp. 551–561.

[48] Shane Malone, Barry Solan, Kieran Collins, and DA Doran. "The metabolic power and energetic demands of elite Gaelic football match play." In: *Journal of Sports Medicine and Physical Fitness* (2016).

[49] Valerio Antonini, Alessandra Mileo, and Mark Roantree. "Engineering features from raw sensor data to analyse player movements during competition". In: *Sensors* 24.4 (2024), p. 1308.

[50] Valerio Antonini, Dermot Sheridan, and Mark Roantree. "A multidimensional database of in-game player movements (Actions and events) in gaelic football". In: *Data in Brief* 57 (2024), p. 111082.

[51] Valerio Antonini, Michael Scriney, Alessandra Mileo, and Mark Roantree. "Using a Spatial Grid Model to Interpret Players Movement in Field Sports". In: *International Conference on Big Data Analytics and Knowledge Discovery*. Springer. 2024, pp. 395–400.

[52] Dermot Sheridan, Valerio Antonini, Michael Scriney, and Mark Roantree. "Automated Detection of Shot Events in Game Phases Using GNSS Data from a Single Team". In: *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer. 2024, pp. 95–105.

[53] Ermanno Rampinini, Aaron J Coutts, Carlo Castagna, R Sassi, and FM Impellizzeri. "Variation in top level soccer match performance". In: *International journal of sports medicine* 28.12 (2007), pp. 1018–1024.

[54] Declan Gamble, Matt Spencer, Andrew McCarren, and Niall Moyna. "Activity profile, PlayerLoad™ and heart rate response of Gaelic football players: A pilot study". In: *Journal of Human Sport and Exercise* (2019).

[55] Aaron J Coutts, John Quinn, Joel Hocking, Carlo Castagna, and Ermanno Rampinini. "Match running performance in elite Australian Rules Football". In: *Journal of Science and Medicine in Sport* 13.5 (2010), pp. 543–548.

[56] Bryan D Cullen, Mark T Roantree, Andrew L McCarren, David T Kelly, Paul L O'Connor, Sarah M Hughes, Pat G Daly, and Niall M Moyna. "Physiological profile and activity pattern of minor Gaelic football players". In: *The Journal of Strength & Conditioning Research* 31.7 (2017), pp. 1811–1820.

[57] Frank Imbach, Waleed Ragheb, Valentin Leveau, Romain Chailan, Robin Candau, and Stephane Perrey. "Using global navigation satellite systems for modeling athletic performances in elite football players". In: *Scientific reports* 12.1 (2022), p. 15229.

[58] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)". In: *Neurocomputing* 307 (2018), pp. 72–77.

[59] Cise Midoglu, Andreas Kjæreng Winther, Matthias Boeker, Susann Dahl Pettersen, Sigurd Pedersen, Nourhan Ragab, Tomas Kupka, Steven A Hicks, Morten Bredsgaard Randers, Ramesh Jain, et al. "A large-scale multivariate soccer athlete health, performance, and position monitoring dataset". In: *Scientific Data* 11.1 (2024), p. 553.

[60] Afzaal Hussain, Muhammad Adeel Zahid, Usama Ahmed, Shahbaz Nazeer, Kashif Zafar, and Abdul Rauf Baig. "Time-series data to refined insights: a feature engineering-driven approach to gym exercise recognition". In: *IEEE Access* 12 (2024), pp. 100343–100354.

[61] Iztok Fister, Luka Lukač, Alen Rajšp, Luka Pečnik, and Dušan Fister. "A minimalistic toolbox for extracting features from sport activity files". In: *2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES)*. IEEE. 2021, pp. 000121–000126.

[62] Alice J Sweeting, Robert J Aughey, Stuart J Cormack, and Stuart Morgan. "Discovering frequently recurring movement sequences in team-sport athlete spatiotemporal data". In: *Journal of Sports Sciences* 35.24 (2017), pp. 2439–2445.

[63] Li Yujian and Liu Bo. "A normalized Levenshtein distance metric". In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1091–1095.

[64] Ryan White, Anna Palczewska, Dan Weaving, Neil Collins, and Ben Jones. "Sequential movement pattern-mining (SMP) in field-based team-sport: A framework for quantifying spatiotemporal data and improve training specificity?" In: *Journal of Sports Sciences* 40.2 (2022), pp. 164–174.

[65] Victor Elijah Adeyemo, Anna Palczewska, and Ben Jones. "Lccspm: l-length closed contiguous sequential patterns mining algorithm to find frequent athlete movement patterns from gps". In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2021, pp. 455–460.

[66] Neil Collins, Ryan White, Anna Palczewska, Dan Weaving, Nicholas Dalton-Barron, and Ben Jones. "Moving beyond velocity derivatives; using global positioning system data to extract sequential movement patterns at different levels of rugby league match-play". In: *European Journal of Sport Science* 23.2 (2023), pp. 201–209.

[67] Victor Elijah Adeyemo, Anna Palczewska, Ben Jones, and Dan Weaving. "Identification of pattern mining algorithm for rugby league players positional groups separation based on movement patterns". In: *Plos one* 19.5 (2024), e0301608.

[68] Dominik Raabe, Reinhard Nabben, and Daniel Memmert. "Graph representations for the analysis of multi-agent spatiotemporal sports data". In: *Applied Intelligence* 53.4 (2023), pp. 3783–3803.

[69] José Gama, Pedro Passos, Keith Davids, Hugo Relvas, João Ribeiro, Vasco Vaz, and Gonçalo Dias. "Network analysis and intra-team activity in attacking phases of professional football". In: *International Journal of Performance Analysis in Sport* 14.3 (2014), pp. 692–708.

[70] Frutuoso GM Silva, Abel JP Gomes, Quoc Trong Nguyen, Fernando ML Martins, and Filipe M Clemente. "A new tool for network analysis on team sports the ultimate performance analysis tool". In: *2017 International Conference*

on Engineering, Technology and Innovation (ICE/ITMC). IEEE. 2017, pp. 439–445.

[71] Bruno Gonçalves, Diogo Coutinho, Sara Santos, Carlos Lago-Penas, Sergio Jiménez, and Jaime Sampaio. "Exploring team passing networks and player movement dynamics in youth association football". In: *PloS one* 12.1 (2017), e0171156.

[72] Gennady Andrienko, Natalia Andrienko, Gabriel Anzer, Pascal Bauer, Guido Budziak, Georg Fuchs, Dirk Hecker, Hendrik Weber, and Stefan Wrobel. "Constructing spaces and times for tactical analysis in football". In: *IEEE Transactions on Visualization and Computer Graphics* 27.4 (2019), pp. 2280–2297.

[73] Ulf Brefeld, Jan Lasek, and Sebastian Mair. "Probabilistic movement models and zones of control". In: *Machine Learning* 108.1 (2019), pp. 127–147.

[74] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. "Large-scale analysis of soccer matches using spatiotemporal tracking data". In: *2014 IEEE international conference on data mining*. IEEE. 2014, pp. 725–730.

[75] Sofia Fonseca, João Milho, Bruno Travassos, and Duarte Araújo. "Spatial dynamics of team sports exposed by Voronoi diagrams". In: *Human movement science* 31.6 (2012), pp. 1652–1659.

[76] Vinicius Machado, Roger Leite, Felipe Moura, Sergio Cunha, Filip Sadlo, and João LD Comba. "Visual soccer match analysis using spatiotemporal positions of players". In: *Computers & Graphics* 68 (2017), pp. 84–95.

[77] Didier A Vega-Oliveros, Moshé Cotacallapa, Leonardo N Ferreira, Marcos G Quiles, Liang Zhao, Elbert EN Macau, and Manoel F Cardoso. "From spatio-temporal data to chronological networks: An application to wildfire analysis". In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019, pp. 675–682.

[78] Leonardo N Ferreira, Didier A Vega-Oliveros, Moshé Cotacallapa, Manoel F Cardoso, Marcos G Quiles, Liang Zhao, and Elbert EN Macau. "Spatiotemporal data analysis with chronological networks". In: *Nature communications* 11.1 (2020), p. 4036.

[79] Jianmin Jia, Chunsheng Liu, Xiaohan Wang, Hui Zhang, and Yan Xiao. "Understanding bike-sharing mobility patterns in response to the COVID-19 pandemic". In: *Cities* 142 (2023), p. 104554.

[80] Pietro Enrico di Prampero, S Fusi, Luigino Sepulcri, Jean-Benoıt Morin, Alain Belli, and Guglielmo Antonutto. "Sprint running: a new energetic approach". In: *Journal of experimental Biology* 208.14 (2005), pp. 2809–2816.

[81] Cristian Osgnach, Stefano Poser, Riccardo Bernardini, Roberto Rinaldo, Pietro Enrico Di Prampero, et al. "Energy cost and metabolic power in elite soccer: a new match analysis approach". In: *Med Sci Sports Exerc* 42.1 (2010), pp. 170–178.

[82] Pietro Enrico di Prampero and Cristian Osgnach. "Metabolic power in team sports-part 1: an update". In: *International journal of sports medicine* 39.08 (2018), pp. 581–587.

[83] Cristian Osgnach and Pietro Enrico di Prampero. "Metabolic power in team sports-Part 2: aerobic and anaerobic energy yields". In: *International journal of sports medicine* 39.08 (2018), pp. 588–595.

[84] Damien Young, Shane Malone, Kieran Collins, Laurent Mourot, Marco Beato, and Giuseppe Coratella. "Metabolic power in hurling with respect to position and halves of match-play". In: *PloS one* 14.12 (2019), e0225947.

[85] Alessio Rossi, Enrico Perri, Luca Pappalardo, Paolo Cintia, Giampietro Alberti, Darcy Norman, and F Marcello Iaia. "Wellness forecasting by external and internal workloads in elite soccer players: a machine learning approach". In: *Frontiers in Physiology* 13 (2022), p. 896928.

[86] Attila Biró, Antonio Ignacio Cuesta-Vargas, and László Szilágyi. "AI-assisted fatigue and stamina control for performance sports on IMU-generated multivariate times series datasets". In: *Sensors* 24.1 (2024), p. 132.

[87] Talko B Dijkhuis, Matthias Kempe, and Koen APM Lemmink. "Early prediction of physical performance in elite soccer matches—a machine learning approach to support substitutions". In: *Entropy* 23.8 (2021), p. 952.

[88] Gianluca Morciano, Andrea Zingoni, and Giuseppe Calabrò. "Optimization and comparison of machine learning algorithms for the prediction of the performance of football players". In: *Neural Computing and Applications* 36.31 (2024), pp. 19653–19666.

[89] Cristian Osgnach, Stefano Poser, Riccardo Bernardini, Roberto Rinaldo, Pietro Enrico Di Prampero, et al. "Energy cost and metabolic power in elite soccer: a new match analysis approach". In: *Med Sci Sports Exerc* 42.1 (2010), pp. 170–178.

[90] Keven Richly, Max Bothe, Tobias Rohloff, and Christian Schwarz. "Recognizing compound events in spatio-temporal football data". In: *International conference on internet of things and big data*. Vol. 2. SCITEPRESS. 2016, pp. 27–35.

[91] Daniel Link and Martin Hoernig. "Individual ball possession in soccer". In: *PloS one* 12.7 (2017), e0179953.

[92] Victor Khaustov and Maxim Mozgovoy. "Recognizing events in spatiotemporal soccer data". In: *Applied Sciences* 10.22 (2020), p. 8046.

[93] Ferran Vidal-Codina, Nicolas Evans, Bahaeddine El Fakir, and Johsan Billingham. "Automatic event detection in football using tracking data". In: *Sports Engineering* 25.1 (2022), p. 18.

[94] Jonas Bischofberger, Arnold Baca, and Erich Schikuta. "Event detection in football: Improving the reliability of match analysis". In: *Plos one* 19.4 (2024), e0298107.

[95] Shane Mangan, Martin Ryan, Simon Devenney, Aidan Shovlin, Jason McGahan, Shane Malone, Cian O'Neill, Con Burns, and Kieran Collins. "The relationship between technical performance indicators and running performance in elite Gaelic football". In: *International Journal of Performance Analysis in Sport* 17.5 (2017), pp. 706–720.

[96] Marco Beato, Giuseppe Coratella, Adam Stiff, and Antonio Dello Iacono. "The validity and between-unit variability of GNSS units (STATSports Apex 10 and 18 Hz) for measuring distance and peak speed in team sports". In: *Frontiers in physiology* 9 (2018), p. 1288.

[97] Marco Beato, Cion Wren, and Kevin L de Keijzer. "The interunit reliability of global navigation satellite systems Apex (STATSports) metrics during a standardized intermittent running activity". In: *The Journal of Strength & Conditioning Research* 38.2 (2024), e49–e55.

[98] James J Malone, Ric Lovell, Matthew C Varley, and Aaron J Coutts. "Unpacking the black box: applications and considerations for using GPS devices in sport". In: *International journal of sports physiology and performance* 12.s2 (2017), S2–18.

[99] Laurence AF Park, Dawn Scott, and Ric Lovell. "Velocity zone classification in elite women's football: where do we draw the lines?" In: *Science and Medicine in Football* 3.1 (2019), pp. 21–28.

[100] Alice J Sweeting, Stuart J Cormack, Stuart Morgan, and Robert J Aughey. "When is a sprint a sprint? A review of the analysis of team-sport athlete activity profile". In: *Frontiers in physiology* 8 (2017), p. 432.

[101] Ben Teune, Carl Woods, Alice Sweeting, Mathew Inness, and Sam Robertson. "A method to inform team sport training activity duration with change point analysis". In: *PLoS One* 17.3 (2022), e0265848.

[102] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[103] Jack Hanson, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. "Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks". In: *Bioinformatics* 33.5 (2017), pp. 685–692.

[104] Zijun Zhang. "Improved adam optimizer for deep neural networks". In: *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee. 2018, pp. 1–2.

[105] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. "1D convolutional neural networks and applications: A survey". In: *Mechanical systems and signal processing* 151 (2021), p. 107398.

[106] Hossein Gholamalinezhad and Hossein Khosravi. "Pooling methods in deep neural networks, a review". In: *arXiv preprint arXiv:2009.07485* (2020).

[107] François Chollet et al. *Keras: Deep learning library for theano and tensorflow*. 2015.

[108] Joachim Gudmundsson and Michael Horton. "Spatio-temporal analysis of team sports". In: *ACM Computing Surveys (CSUR)* 50.2 (2017), pp. 1–34.

[109] Ulrik Brandes. "A faster algorithm for betweenness centrality". In: *Journal of mathematical sociology* 25.2 (2001), pp. 163–177.

[110] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[111] Siavash Saki, Jonas Hamann, and Tobias Hagen. "TessPy: a python package for geographical tessellation". In: *Journal of Open Source Software* 7.76 (2022), p. 4620.

[112] Marc Barthelemy. "Betweenness centrality in large complex networks". In: *The European physical journal B* 38.2 (2004), pp. 163–168.

[113] Greg Doncaster, Richard Page, Paul White, Robert Svenson, and Craig Twist. "Analysis of physical demands during youth soccer match-play: Considerations of sampling method and epoch length". In: *Research quarterly for exercise and sport* 91.2 (2020), pp. 326–334.

[114] Zunqi Niu, Jingyun Yang, Jiewei Long, Xinyi Zhang, Zhaoyang Wang, Iyán Iván-Baragaño, and Hongyou Liu. "A new lens on football match intensity: from the most demanding passage to extremely demanding passage". In: *BMC Sports Science, Medicine and Rehabilitation* 17.1 (2025), p. 290.

[115] Liam D Harper, Daniel J West, Emma Stevenson, and Mark Russell. "Technical performance reduces during the extra-time period of professional soccer match-play". In: *PloS one* 9.10 (2014), e110995.

[116] Bradley Thoseby, Andrew D Govus, Anthea Clarke, Kane J Middleton, and Ben Dascombe. "Positional and temporal differences in peak match running demands of elite football". In: *Biology of Sport* 40.1 (2023), pp. 311–319.

[117] Pietro E Di Prampero, Alberto Botter, and Cristian Osgnach. "The energy cost of sprint running and the role of metabolic power in setting top performances". In: *European journal of applied physiology* 115 (2015), pp. 451–469.

[118] Paulo Gaudino, FM Iaia, Giampietro Alberti, AJ Strudwick, Greg Atkinson, and Warren Gregson. "Monitoring training in elite soccer players: systematic bias between running speed and metabolic power data". In: *International journal of sports medicine* 34.11 (2013), pp. 963–968.

[119] Adrian J Gray, Kathleen Shorter, Cloe Cummins, Aron Murphy, and Mark Waldron. "Modelling movement energetics using global positioning system devices in contact team sports: limitations and solutions". In: *Sports Medicine* 48 (2018), pp. 1357–1368.

[120] Ted Polglaze and Matthias W Hoppe. "Metabolic power: A step in the right direction for team sports". In: *International journal of sports physiology and performance* 14.3 (2019), pp. 407–411.

[121] Eoghan Boyle, Shane Malone, Shane Mangan, Dominic A Doran, and Kieran Collins. "Applied sport science of Gaelic football". In: *Sport Sciences for Health* 19.1 (2023), pp. 27–39.

[122] Sarah da Glória Teles Bredt, Mauro Heleno Chagas, Gustavo Henrique Peixoto, Hans Joachim Menzel, and André Gustavo Pereira de Andrade. "Understanding player load: Meanings and limitations". In: *Journal of human kinetics* 71 (2020), p. 5.

[123] Howard Chen, Mark C Schall, and Nathan B Fethke. "Gyroscope vector magnitude: A proposed method for measuring angular velocities". In: *Applied ergonomics* 109 (2023), p. 103981.

[124] D Augusto, J Brito, R Aquino, et al. *Contextual variables affect peak running performance in elite soccer players: A brief report. Frontiers in Sports and Active Living. 2022; 4.*

[125] Nicolas Le Page and Daniel Link. "Load peak intensity in pre-season training sessions and official competitions in professional soccer". In: *International Journal of Sports Science & Coaching* 20.2 (2025), pp. 669–678.

[126] Daniel J Cunningham, David A Shearer, Neil Carter, Scott Drawer, Ben Pollard, Mark Bennett, Robin Eager, Christian J Cook, John Farrell, Mark Russell, et al. "Assessing worst case scenarios in movement demands derived from global positioning systems during international rugby union matches: Rolling averages versus fixed length epochs". In: *PloS one* 13.4 (2018), e0195197.

[127] Ronald Yurko, Samuel Ventura, and Maksim Horowitz. "nflWAR: a reproducible method for offensive player evaluation in football". In: *Journal of Quantitative Analysis in Sports* 15.3 (2019), pp. 163–183.

[128] Andrey V Skripnikov, Ahmet Cemek, and David Gillman. "Leveraging minute-by-minute soccer match event data to adjust Team's offensive production for game context". In: *Journal of Quantitative Analysis in Sports* 0 (2025).

[129] Muhammad Asif and Ian G McHale. "In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model". In: *International Journal of Forecasting* 32.1 (2016), pp. 34–43.

[130] Alice J Sweeting123, Stuart Morgan, Stuart Cormack, and Robert J Aughey. "A movement sequencing analysis of team-sport athlete match activity profile". In: (2014).

[131] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[132] Hongchun Qu and Zheng Zhang. "A time series data augmentation method based on SMOTE". In: *2024 36th Chinese Control and Decision Conference (CCDC)*. IEEE. 2024, pp. 5336–5341.

[133] Hong Cao, Xiao-Li Li, David Yew-Kwong Woon, and See-Kiong Ng. "Integrated oversampling for imbalanced time series classification". In: *IEEE Transactions on Knowledge and Data Engineering* 25.12 (2013), pp. 2809–2822.

[134] Chaomin Wu, Xu Cheng, and Hao Wang. "A contrastive clustering loss function increases class-balanced in time series classification". In: *Expert Systems with Applications* 283 (2025), p. 127493.

[135] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[136] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[137] Lyndell Bruce, Dan Dwyer, and Aaron Fox. "Reliability of live and video-based coding in netball using the NetballStats application". In: *Plos one* 17.6 (2022), e0269330.

[138] Declan Gamble, Jonathan Bradley, Andrew McCarren, and Niall M Moyna. "Team performance indicators which differentiate between winning and losing in elite Gaelic football". In: *International Journal of Performance Analysis in Sport* 19.4 (2019), pp. 478–490.

[139] Maike Stoeve, Dominik Schuldhaus, Axel Gamp, Constantin Zwick, and Bjoern M Eskofier. "From the laboratory to the field: IMU-based shot and pass detection in football training and game scenarios using deep learning". In: *Sensors* 21.9 (2021), p. 3071.

[140] Puneet Misra and Arun Singh Yadav. "Improving the classification accuracy using recursive feature elimination with cross-validation". In: *Int. J. Emerg. Technol* 11.3 (2020), pp. 659–665.

[141] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1 (2002), pp. 389–422.

[142] Arif Mudi Priyatno, Triyanna Widiyaningtyas, et al. "A systematic literature review: recursive feature elimination algorithms". In: *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)* 9.2 (2024), pp. 196–207.

[143] Serpil Ustebay, Zeynep Turgut, and Muhammed Ali Aydin. "Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier". In: *2018 international congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT)*. IEEE. 2018, pp. 71–76.

[144] Bharat Richhariya, Muhammad Tanveer, Ashraf Haroon Rashid, Alzheimer's Disease Neuroimaging Initiative, et al. "Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination

(USVM-RFE)". In: *Biomedical Signal Processing and Control* 59 (2020), p. 101903.

[145] Adi Zaenul Mustaqim, Sumarni Adi, Yoga Pristyanto, and Yuli Astuti. "The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection". In: *2021 International conference on artificial intelligence and computer science technology (ICAICST)*. IEEE. 2021, pp. 270–275.

[146] Andrew Y Ng. "Feature selection, L 1 vs. L 2 regularization, and rotational invariance". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 78.

[147] David R Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 20.2 (1958), pp. 215–232.

[148] Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[149] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[150] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 20–29.