Instrumented Mouthguards: Advancing the Functionality and Reliability

David Rhys Lloyd Powell

Submitted to Swansea University in fulfilment of the requirements for the degree of Doctor of Philosophy

Swansea University

2024

Abstract

Instrumented mouthguards (IMGs) are wearable devices designed to record kinematic data describing the head's motion during potentially injurious impacts in sports. Already a popular device, IMGs are set for a surge in popularity following World Rugby's mandate making them a requirement for all professional rugby players. Despite this widespread implementation, numerous aspects of the design may be improved. Specific improvements include stricter measures to ensure the validity of recorded data and the extraction of further information about the recorded data while avoiding time-consuming video review. Machine learning algorithms have been created to address the validation issue, but not for female sports specifically or rugby union. To address this, a dataset was collected from six women's rugby union matches, resulting in 214 impacts from 480 minutes of play. After training, a machine learning algorithm yielded scores of 0.92 and 0.85 for the area under the receiver operator and precision-recall curves (AUROC/AUPRC) respectively, on test data. This advancement signifies a crucial step in female sports' head impact telemetry, enhancing safety and data reliability in contact sports for women. Secondly, a study used kinematic recordings to create algorithms predicting impact action (ball carrier vs. tackler) and impact type (direct head contact vs. secondary acceleration). Machine learning algorithms achieved 69.4%/0.721 and 65.4%/0.744 macro recall/AUROC scores. With further refinement, this may potentially automate impact analysis, aiding athlete protection. Lastly, methods to reliably report linear acceleration are not fully understood. This was investigated by estimating the linear acceleration at the head's centre of gravity with measurements from a cohort of 25 (11F) individuals. Substantial differences between maximum and minimum impact values were found. Given the variation in head shape and size between youth, adult males and adult females, this indicates a one-size-fits-all approach will not be appropriate and individualised measurements are required to estimate linear acceleration accurately.

Declarations

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed....[redacted].....

Date....02/12/2024.....

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed.....[redacted].....

Date....02/12/2024.....

I hereby give consent for my thesis, if accepted, to be available for electronic sharing.

Signed.....[redacted].....

Date....02/12/2024.....

The University's ethical procedures have been followed and, where appropriate, ethical approval has been granted.

Signed....[redacted].....

Date....02/12/2024.....

Table of Contents

Acknowledgements

I would also like to pass my sincere gratitude to my supervisors, Dr Elisabeth Williams, Associate Professor Hari Arora, Professor Paul Docherty, and Dr Desney Greybe. Dr Elisabeth Williams' knowledge, effort, and enthusiasm have not only made this work possible but enjoyable as well. I could not have hoped for a better primary supervisor. I am thankful for the time and effort provided by Associate Professor Hari Arora, whose thoughtful suggestions and support significantly shaped this thesis. Professor Paul Docherty's analytical approach and work ethic have not only helped to improve the work within this thesis but have also helped me to become a better researcher myself. The guidance of Dr Desney Greybe allowed me to find my feet when first undertaking this project, providing me with foundational skills and knowledge that are now fundamental to the thesis.

To even reach the point of undertaking a PhD requires massive investment from those around you. Mum, Dad, and Siân, your support over the years has allowed me the opportunity to reach this point, I truly could not have done it without you. To Mr William Morgan and my teachers at The Cotswold School, thank you for the faith you put in me during the sixth form, without it I would not be where I am today. Finally, to Jenny, you've always been there for me when I needed it and for that, I am forever grateful.

List of Tables

List of Figures

Abbreviations

ATD	Anthropometric Test Device
AUPRC	Area Under the Precision Recall Curve
AUROC	Area Under the Receiver Operator Curve
BC	Ball Carrier
BUCS	British Universities and Colleges Sport
CG	Centre of Gravity
CTE	Chronic Traumatic Encephalopathy
CHAMP	Consensus Head Acceleration Measurement Practices
CCC	Concordance Correlation Coefficient
CWT	Continuous Wavelet Transformation
DFT	Discrete Fourier Transformation
DHC	Direct Head Contact
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FCQ	Frequency Correlation Quotient
HIT	Head Impact Telemetry
IMU	Inertial Measurement Unit
IMG	Instrumented Mouth Guard
mTBI	Mild Traumatic Brain Injury
ML	Machine Learning
MLP	Multilayer Perceptron
NDC	No Direct Head Contact
NFL	National Football League
OOP	Object Orientated Programming
PPE	Personal Protective Equipment
PPV	Positive Predictive Value
PSD	Power Spectral Density
RFU	Rugby Football Union
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine

TBI	Traumatic Brain Injury
TN	True Negative
TP	True Positive
WRU	Welsh Rugby Union

"The human brain has 100 billion neurons, each neuron connected to 10,000 other neurons. Sitting on your shoulders is the most complicated object in the known universe." - Dr Michio Kaku

1 Introduction

Recent years have proved a tumultuous time for sports with the emergence of the concussion crises. Rugby union (RU) is one of the affected sports, with many players now diagnosed with chronic neurodegenerative disease as a result of their playing careers. The use of Instrumented mouthguards (IMGs) has been mandated to assist with concussion detection by World Rugby to better protect athletes' health. This thesis focuses on advancing the reliability and functionality of IMGs through the application of machine learning (ML) and other innovative approaches. This introductory chapter will cover the background to sports concussion crisis and offer a brief overview of the key themes of the thesis, which are discussed in greater depth in later chapters.

Sections 1.1 and 0 describe the terminology relating to brain injury and an overview of the pathophysiology of brain injury. Section 1.3 overviews the history of brain injury research in sports, leading to the present day. Section 0 discusses the rates of injury in modern RU, while 1.5 introduces IMGs and World Rugby's use of IMGs. Section 1.6 provides a brief introduction to ML, while 0 and 1.8 provide the thesis aims and outline respectively.

1.1 Brain Injury Terminology

This thesis will commonly use the following definitions regarding brain injury.

Chronic Traumatic Encephalopathy (CTE) is a neurodegenerative disease associated with repeated head impacts, sometimes seen in athletes and military personnel (Gavett et al., 2011; McKee, 2020; Omalu et al., 2005; Stein et al., 2014). It shares some neuropathological features with Alzheimer's disease, including tau pathology, but is considered a distinct entity (Turner et al., 2016). CTE can lead to early-onset dementia, with symptoms including mood changes, cognitive impairment, and motor dysfunction (Montenigro et al., 2015).

Sub-concussion refers to brain damage that occurs without overt clinical symptoms, making it particularly insidious and challenging to detect (Bailes et al., 2013; Nowinski et al., 2022). Although sub-concussive impacts lack immediate signs of impairment, evidence suggests that repeated exposure to such impacts can have significant long-term consequences, contributing to neurodegenerative conditions like

early-onset dementia and CTE (McKee, 2020; Nowinski et al., 2022). Studies on athletes in contact sports and military personnel exposed to repetitive head trauma underscore the cumulative effects of sub-concussions, revealing associations with structural brain changes, cognitive decline, and behavioural disturbances over time (Breedlove et al., 2012; Mez et al., 2016). This emerging understanding highlights the critical need for strategies to monitor and mitigate sub-concussive impacts, even in the absence of overt injury symptoms, to prevent progressive neurological damage.

Concussion, or mild traumatic brain injury (mTBI), is a less severe form of brain injury that, despite its classification, can have lasting effects on cognitive health, emotional well-being, and overall quality of life, particularly when complicated by repeated injuries or inadequate recovery (Carroll et al., 2004). Concussions frequently result from sports-related impacts, falls, or motor vehicle accidents and present with a range of symptoms, including headaches, dizziness, sensitivity to light or sound, memory difficulties, and transient confusion or disorientation (Giza & Hovda, 2001; Patricios et al., 2023). While many individuals recover within weeks, some experience prolonged symptoms, referred to as post-concussion syndrome, which can persist for months or longer, highlighting the potential for significant disruption to daily life (Daneshvar et al., 2011; Gardner & Yaffe, 2015). Like sub-concussive injuries, an accumulation of mTBIs may contribute to neurodegenerative conditions (Gavett et al., 2011; Nowinski et al., 2022).

Traumatic brain injury (TBI) broadly refers to brain damage resulting from an external force that disrupts normal brain function, ranging in severity from mild to severe (Menon et al., 2010; Saatman et al., 2008). Severe TBIs are characterized by extended periods of unconsciousness, significant memory deficits, and persistent physical and cognitive impairments, often necessitating long-term medical intervention and rehabilitation (Maas et al., 2017; Saatman et al., 2008). Such injuries can result from high-impact events such as motor vehicle accidents, falls, or violent trauma and are associated with profound social and economic burdens, including reduced quality of life for survivors and substantial healthcare costs (Roozenbeek et al., 2013). Recovery depends on injury severity, individual characteristics, and access to medical services (Pervez et al., 2018). While most mTBI patients recover within months, some experience persistent symptoms (Azouvi et al., 2017).

The following section provides a deeper look at the pathophysiology of these brain injuries.

1.2 Brain Injury Pathophysiology

Brain injuries are highly complex, involving intricate biochemical and cellular processes, and a comprehensive exploration of this topic is beyond the scope of this thesis. However, a brief overview of the pathophysiology relevant to this work will be provided here, with more detailed explanations available in the works of Capizzi et al. (2020) and Ng & Lee (2019).

Primary injury represents the immediate physical damage to brain tissue following traumatic brain injury (TBI), resulting directly from mechanical forces applied during the initial impact (Giza & Hovda, 2001; S. Y. Ng & Lee, 2019). Primary injury is classified into focal and diffuse types, with diffuse injuries occurring more frequently in TBIs. Focal injuries are localized and often result from direct head impacts, leading to contusions, brain lacerations, and haemorrhage in specific brain regions (Capizzi et al., 2020; S. Y. Ng & Lee, 2019). Diffuse injury, on the other hand, typically results from rapid brain acceleration without direct contact, causing widespread shearing and stretching of neuronal and vascular tissues, including axons and oligodendrocytes, particularly within the brainstem and corpus callosum (Meythaler et al., 2001). This type of injury is characterized by axonal injury throughout the brain, which is a key determinant of TBI severity (Fujita et al., 2012; Gennarelli et al., 1987).

Secondary injury occurs as a delayed response to the primary injury, potentially lasting from hours to years and involving various neurochemical and metabolic disruptions (Loane & Faden, 2010; S. Y. Ng & Lee, 2019). Key mechanisms of secondary injury include excitotoxicity and mitochondrial dysfunction, both of which exacerbate neural damage over time. Excitotoxicity involves the breakdown of the blood-brain barrier, leading to the excessive release of excitatory neurotransmitters, such as glutamate, which can cause oxidative stress and neuronal cell death (Loane & Faden, 2010; S. Y. Ng & Lee, 2019). Mitochondrial dysfunction results from disrupted calcium homeostasis and excess ions entering cells post-injury, leading to mitochondrial membrane depolarisation, ATP synthesis inhibition, and reduced cellular energy for repair (Xiong et al., 1997).

Following this brief description of brain injury, Section 1.3 describes how the understanding of these injuries in sport has developed, from the 1920's until now.

1.3 A Brief History of Brain Injury Research in Sport

The origins of sports-related brain injury research are attributed to the work of Dr W. Trotter, who observed brain injuries in boxers in the 1920s (Changa et al., 2018). In this transcribed oration session entitled "Certain minor injuries to the brain", he described concussion in athletes as a transient state, followed by amnesia and the absence of clear structural cerebral injury (Trotter, 1924). As the field developed, publications from Dr Harrison Martland, Dr C. B. Cassasa, Dr Michael Osnato, and Dr Vincent Gilberti provided further evidence in support of Trotter's findings (Changa et al., 2018). These authors reported what was believed to be physical evidence of concussion, noting 'bundles of fibrils' and haemorrhages, leading to conclusions that concussions may also have chronic effects (Changa et al., 2018; Martland, 1928). This growing body of evidence highlighting links between concussions and chronic effects led Dr Martland to describe the classical clinical case of the "Punch Drunk" boxer (Changa et al., 2018; Martland, 1928).

The punch drunk syndrome includes acute, delayed, and chronic symptoms of brain injury, ranging from changes in gait between rounds to Parkinsonian gait, tremulousness, and cognitive decline (Changa et al., 2018). It reported that boxers who had a higher head impact exposure during their career would be more likely to exhibit these symptoms, in particular, low-skill boxers who were knocked down several times a day in sparring practice (Changa et al., 2018; Martland, 1928). In the years after their careers, former boxers saw premature dementia and disorientation, suspected to be a result of repeated brain injury (Corsellis et al., 1973; Milspaugh, 1937).

"Punch drunk" or "dementia pugilistica" was believed to be a disease only found in boxers, until an autopsy of a former National Football League (NFL) athlete revealed that they had suffered from the disease (Omalu et al., 2005). The pre-mortem medical history described many of the symptoms from the studies of the early 20th century, suspected to be the result of head trauma from their NFL career (Omalu et al., 2005). The disease now called CTE, was later found in 86% of American football players who donated their brains to a brain bank (Mez et al., 2016). This led to a prolonged legal battle, with former players suing the NFL for masking the severity of repeated

head trauma, culminating in a reported settlement of $1 billion (Ventresca & Henne, 2020).

CTE has been diagnosed in former professional athletes from RU, ice hockey, mixed martial arts, and other high-impact sports (McKee et al., 2009; Mez et al., 2016; Montenigro et al., 2015; Nowinski et al., 2022). These discoveries and the increased awareness of concussions led governing bodies to develop targeted preventive strategies (Arbogast et al., 2022; Koerte et al., 2021; McNamee et al., 2023; Patricios et al., 2023). These preventive strategies include the introduction of concussion protocols, mandatory rest periods, and stricter return-to-play guidelines for athletes, especially in high-impact sports such as football, rugby, and hockey (Raftery & Falvey, 2022). It has also seen the introduction of touchline assessments, like the Sports Concussion Assessment Tool (SCAT), which have also been standardised across many sports to facilitate early detection and improve injury management (Stemper et al., 2019). Additionally, rule changes, such as limiting high tackles in rugby and reducing contact drills in football practices, aim to reduce the frequency and intensity of head impacts (Patricios et al., 2018).

The growing recognition and/or detection of concussions have driven a substantial increase in scientific research focusing on sports-related head trauma. This field encompasses diverse approaches, including the use of wearable technologies for real-time monitoring, medical imaging for injury assessment, simulation-based studies for understanding impact dynamics, and large-scale epidemiological investigations (Arbogast et al., 2022; Giza & Hovda, 2001; Patricios et al., 2018). Collectively, these efforts aim to quantify the prevalence and severity of brain injuries, reduce their occurrence and impact, and unravel the complex physiological mechanisms underlying concussions and their long-term effects (Giza & Hovda, 2001; Patricios et al., 2023).

Despite the research and interventions, RU is now seeing similar legal battles to that of the NFL, which has garnered significant attention from media outlets and academia alike (Ingle, 2023). Players who have been diagnosed with neurological issues, including early onset dementia, are seeking damages from World Rugby, the Rugby Football Union (RFU), and the Welsh Rugby Union (WRU), alleging negligence and inadequate player protection (Ingle, 2023). As reported in mainstream media, the case

currently involves 268 players, with an additional 27 players recently joining the proceedings (Davies, 2023; Ingle, 2023).

This illustrates the problem concussions have been and continue to be in sports. The next stages of this chapter will focus on the specific issue that concussion plays RU and investigate how wearable technology, specifically instrumented mouthguards, is being used to combat it.

1.4 The Rate of Brain Injury in Rugby Union

The frequency of brain injuries in sports has now been documented in varying sports, age groups, and levels of competition (Dompier et al., 2015; Stokes, Kemp, et al., 2023). In keeping with the themes of this thesis, only interventions conducted in RU will be assessed in this section to provide an overview of the available research. In general, for a concussion to be reported within these studies it must be diagnosed by a medical professional. Whilst this aims to provide consistency between studies, this requirement may make reporting accurate figures from community sport more difficult, due to reduced access to medical staff. Self-reported concussions may also be unreliable as concussions are underreported by athletes for various reasons (Fraas et al., 2014).

RU's injury rates are well reported in England, as the governing body, the Rugby Football Union (RFU), has provided an injury audit since 2002 (West et al., 2021). Whilst this initially focused on the male professional game, it now reports injuries for males and females, in youth, community, university, and professional RU (West et al., 2021). These RFU injury audits have reported concussion to be the most prevalent injury in the male professional game since the 2011/12 rugby season (K. Stokes, Kemp, et al., 2023). Similar findings have come from the Irish and Welsh professionals, with Irish professionals experiencing rates of incidence rate of 18.4 diagnosed concussions /1000 hours (Cosgrave & Williams, 2019), and the Welsh counterparts 21.4/1000hr in the 15/16 season (Moore et al., 2015). Further to this, the injury rates reported within the men's rugby World Cup injury audits have illustrated similar trends and values (Fuller et al., 2008, 2013, 2016, 2020). Concussion was found to be the most common injury that occurred in matches at all other levels of the game, in both males and females (K. Stokes, Kemp, et al., 2023). Concussions accounted for between 20% and 34% of the total injuries recorded at each level in the

most recent report of the 2021/22 season, with the rates per 1,000 match hours shown in Table 1-1.

Within the female game, concussion is still a common injury, although less research has been available. More studies are now being published, perhaps due to calls for more research in response to the growth of the female game (Palmer & Hargreaves, 2023). A 2022 survey collated published research articles to assess injury rates in women's RU, rugby league, and sevens (King et al., 2022). The pooled rate of concussion in RU from the nine studies was reported as 2.8/1,000 PMH, whilst rugby sevens and league produced rates of 8.9 and 10.3/1,000 PMH respectively (King et al., 2022). The values reported in this study for RU concussion rates were much lower than the rates the RFU have recently reported, Table 1-1, and lower than the average across the 2011/2012–2013/2014 and 2017/2018–2019/2020 seasons (5.0/1,000 PMH) (Starling et al., 2023).

Table 1-1: The frequency of concussion at various levels of rugby union within England.

Level of Play	Injury Rate / 1,000 player match hours	% of total injury
Women's Community (Roberts et al., 2023)	7.0	22
Women's Professional (Williams et al., 2024)	15.0	34
Men's BUCS Super Rugby (Kemp et al., 2022)	17.3	20
Men's Professional Stokes, Kemp, et al., 2023)	18.2	24
Men's Community (Roberts et al., 2023)	4.9	22
Male Youth Rugby Stokes, Roberts, et al., 2023)	8.0	29

Given these incidence rates have failed to decline following the introduction of protocols and interventions, it suggests that more changes must be made to make RU safer to participate in. The next section of this chapter explores how World Rugby, the governing body overseeing RU, plans on using IMGs to help manage the concussion crisis.

1.5 World Rugby's Use of Instrumented Mouthguards for Concussion Management

With the awareness of the concussion crisis, there has been a growing focus on prioritising player safety and the importance of precise injury detection and assessment (Raftery & Falvey, 2022; Tooby et al., 2023). The use of IMGs may aid concussion detection and assessment as it enables the remote measurement of head kinematics during head acceleration events (HAEs) (Greybe et al., 2020; E. M. P. Williams et al., 2021). These mouthguards are equipped with sensors which capture data describing the motion of the head during impacts experienced during gameplay. By capturing this kinematic data, the device offers a promising avenue for better understanding and mitigating the risk of concussions in the sport. They, along with other devices, have now been featured in many academic publications, although questions remain about their reliability (Wu et al., 2018). More information on IMGs is in Chapter 2.

As of January 1, 2024, World Rugby, the international governing body for RU, has mandated the use of IMGs for professional teams worldwide (World Rugby, 2023). This follows the deployment of these devices during the Rugby World Cup 2021 – Women (held in 2022) and the Otago Community Head Impact Detection study (Bussey et al., 2023; World Rugby, 2022). This decision carries profound implications for the research conducted in this doctoral thesis. This mandate reflects a growing recognition of the potential of this technology to enhance injury prevention strategies, refine training methodologies, and promote evidence-based decision-making (Tooby et al., 2023). Whilst they will be initially used to aid injury detection, as IMGs are not currently capable of detecting concussion, the long-term objectives have not been fully defined by World Rugby. Consequently, this highlights the importance of the work in this thesis, which seeks to improve the quality and interpretation of IMG data for contact sports (Palmer & Hargreaves, 2023).

1.6 What is Machine learning and how has it Supported the Use of Instrumented Mouthguards

In ML, algorithms play a central role in building models that can learn patterns from data and make predictions or decisions (Mitchell, 1997). An algorithm is defined by Cormen et al., 2009 as the following. It is a finite, well-defined sequence of steps or instructions used to solve a problem or perform a computation. It is a methodical process for transforming inputs into outputs by following a logical progression of operations. Algorithms are foundational to computer science and are used across various domains to automate tasks, optimize processes, and analyse data.

ML is a specialized domain within artificial intelligence (AI) that focuses on designing algorithms capable of replicating human decision-making processes and improving performance with experience (Goodfellow et al., 2016). These algorithms not only mimic human-like decisions but also analyse data to adapt and learn patterns autonomously (Mitchell, 1997). A well-known definition of machine learning is as follows: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E” (Mitchell, 1997, p.470).

Classification and regression are among the most widely utilized ML approaches (Alpaydın, 2013). Classification algorithms assign input data to predefined categories, such as predicting whether an email is spam or not, while regression algorithms focus on predicting continuous outcomes, like stock prices or temperature changes (Hastie et al., 2009b). These approaches demonstrate ML’s capacity to address both discrete and continuous data problems, forming a foundation for numerous real-world applications across fields like medicine, finance, and sports analytics (Csizmadia et al., 2022; Janiesch et al., 2021; Shalev-Shwartz & Ben-David, 2013).

ML algorithms can also be categorised based on their learning paradigms, with the primary approaches being supervised, unsupervised, and reinforcement learning (Mitchell, 1997). Supervised learning involves training algorithms on labelled datasets where the input-output relationship is clearly defined, enabling performance evaluation during training (Hastie et al., 2009a). Examples include linear regression, which models numerical relationships, and neural networks, which simulate brain-like structures for advanced pattern recognition (Goodfellow et al., 2016). Unsupervised

learning works with unlabelled data to uncover hidden patterns or structures (Shalev-Shwartz & Ben-David, 2013). Techniques such as k-means clustering and autoencoders are widely used to identify groupings or reduce data dimensions without explicit guidance (Hastie et al., 2009a; Murphy, 2012). Reinforcement learning focuses on sequential decision-making, by interacting with their environment and receiving feedback in the form of rewards or penalties (Sutton & Barto, 2018).

IMGs may report spurious data as genuine, therefore methods must be in place to reduce these occurrences (Kuo, Wu, Hammoor, et al., 2016; Luke et al., 2024; Patton, 2016). Supervised classification algorithms have proven instrumental in advancing the detection, classification, and analysis of HAEs captured by IMGs in numerous sports (Goodin et al., 2021; Raymond et al., 2022; Wu et al., 2014). These studies have used a variety of methods, including different algorithm types, data preparation methods, and the inclusion of synthetic data to improve the performance of the algorithms. The result of this work is the development of classification algorithms with near-human-level performance in the detection of genuine and spurious events, in specific situations (Goodin et al., 2021; Raymond et al., 2022; Wu et al., 2014). These developments underline the role of ML in improving the capabilities of IMGs, ultimately contributing to improved player safety and a better understanding of impact biomechanics.

1.7 Thesis Aims

The data generated by IMGs following World Rugby's mandate offers an unprecedented opportunity to progress concussion research in RU. Not only could this aid the understanding of how the brain responds to impacts, but the interventions it will create will make the sport safer for the 8.5 million players participating in RU worldwide (Salmon et al., 2019). To translate this influx of data into meaningful advances in player safety, it is essential to establish robust methods for data processing and analysis. Accurate and reliable data collection is critical as inaccuracies may invalidate findings and hinder the development of effective interventions. Additionally, data alone will not provide any solutions to the concussion crisis, by developing an effective analysis pipeline researchers will be able to develop insights into the causes and outcomes of brain injury. Therefore, the work within this thesis will present methodologies and findings that look to facilitate accurate, reliable, and insightful work within this field.

This mandate also offers the opportunity to begin addressing the gender disparity in RU injury research, as female athletes remain underrepresented (Palmer & Hargreaves, 2023). This thesis will address this issue by developing targeted solutions for the issues present with IMG use in female RU. Importantly, the research aligns with broader societal conversations about gender equity in data collection, with work from this thesis and the wider research group highlighted in a recent BBC report on the gender data gap (Palmer & Hargreaves, 2023). Such efforts aim to improve player safety, and foster a more inclusive understanding of injury risks.

The first research question focuses on the detection of genuine HAEs in women's RU. This inquiry stems from earlier studies that tackled similar challenges in other sports. For instance, ML has been employed to detect direct head impacts in American football and HAEs in Australian Rules football, predominantly using male-derived datasets (Goodin et al., 2021; Wu et al., 2018). These studies provide valuable insights into classification algorithms, feature sets, and key feature groups. By adapting and enhancing these techniques, this research aims to develop an optimised combination of ML algorithms and features tailored for the detection of HAEs in women's RU. This approach addresses a significant gap, as no existing classifier currently targets either women's sports or RU specifically. Moreover, the creation of such an algorithm will enable a deeper exploration of the unique characteristics of female head acceleration events, offering a foundation for future research and potentially improving safety protocols within the sport.

Currently, IMGs are primarily used for impact detection and the estimation of the head's kinematics, providing limited functionality beyond this. Detailed analysis of specific HAEs typically requires a supplementary and resource-intensive video review process. ML offers a potential solution by enabling the automated detection of critical HAE characteristics, reducing the reliance on manual review. Activity recognition, which has demonstrated success across various domains using diverse sensors, presents a promising avenue for application in HIT systems (Dang et al., 2020). By integrating these methodologies, IMG systems could be significantly enhanced, offering time savings and actionable insights into player safety. To evaluate these possibilities, clearly defined research goals are needed. For instance, prior studies have successfully recognised sport-specific actions, making RU motions an intriguing test case for comparison (Hendry et al., 2020; Kautz et al., 2017; Tabrizi et al., 2020). A

key question thus emerges: Can ML differentiate between head accelerations recorded from ball carriers (BCs) and those from tacklers? Building on the ability to characterise head accelerations and identify spurious recordings, the research could progress to addressing a more nuanced question: Can ML distinguish between head accelerations caused by direct head contact and those resulting from indirect or non-contact events? Successfully answering these questions would enhance HIT systems' diagnostic capabilities and contribute to a more sophisticated understanding of head impacts in RU.

A critical aspect of all scientific research is to ensure that results are as reliable and accurate as feasibly possible. A significant concern in this field is the method used to report linear acceleration. A 2022 consensus statement emphasised that linear acceleration should be measured and reported from the head's centre of gravity (CG) to improve reliability (Arbogast et al., 2022). However, this recommendation has yet to be explored in detail, particularly regarding the potential inaccuracies introduced by varying head sizes across athletes of different ages, sexes, and physical builds. Bussey et al., 2022 highlighted that merely reporting linear acceleration without considering biomechanical nuances, such as individual anatomical differences, is insufficient, as it risks misinterpreting impact severity and biomechanics. This underscores the need for standardised methods that account for these variations to improve data fidelity and its subsequent interpretation. Given that IMGs are used by diverse populations, spanning male and female athletes from youth to professional levels, the research question emerges: What methodologies should be adopted to most effectively translate and report linear acceleration data to ensure accuracy across diverse user groups? Addressing this question is critical to advancing IMG technology and ensuring that data collected contributes meaningfully to safety and injury prevention research in sports.

1.8 Thesis Outline

Chapter 2: A review of the use and validation of instrumented mouthguards

The chapter begins with a brief overview of HIT (HIT) systems in contact sports, describing the purpose and utilisation of HIT systems. A detailed discussion of IMGs follows, including their significance in sports, fundamental features, and design considerations. The device performance is then discussed, with results obtained from

both lab and on-field verifications. Lastly, the methods used to ensure the accuracy and reliability of IMGs are described.

Chapter 3: Important Methodologies

This chapter delves into the computational methodologies used within this field, with a more detailed description of the methods most relevant to this thesis. With the code supporting this thesis written in Python, this chapter begins by exploring the benefits of Python and object-oriented programming. The chapter then moves to provide an overview of action recognition and artificial intelligence, two of the key methodological themes within this thesis. This leads to the description from the first principles of the methods that underpin these concepts, ranging from feature selection and generation to ML classification algorithms.

Chapter 4: Moving from Review to Methods and Results

This chapter provides an interlude to mark the end of the introductory sections of the thesis. A brief describe the layout of the remaining chapters of the thesis is described within.

Chapter 5: Collection, Processing, and Validation of Head Impact Data

Having reviewed the methodologies in prior chapters, Chapter 5 offers the first look at the practical study within this thesis. This chapter reports the collection of data for use in later studies, describing the deployment of the devices, the collection of kinematic and video data, and the cohort from which data were collected. After the collection is reported the methods used to validate and the data are outlined. A discussion follows, consolidating the methodologies, results, and their implications within the context of the study's objectives and the broader landscape of head impact data collection and processing.

Chapter 6: Feature Development and Extraction in Python

Here is outlined the process following the initial data collection and preparation of transforming the recorded data into feature vectors suitable for training ML algorithms. This process encompasses five feature groups: pulse parameters, positional derivatives, power spectral density, wavelet transformation, and other features, as described in Chapter 3. The section further explains the methodology of

feature generation, storage, naming conventions and normalisation. The discussion provides insights into feature selection's impact on the different classification tasks explored within this thesis, emphasising the challenges and considerations in feature selection and classification tasks within the domain.

Chapter 7: Development of a Head Acceleration Classifier

This chapter delves into a study of the creation of ML algorithms, trained to discern between recordings of HAEs and spurious events in women's collegiate rugby. With prior research conducted within this field, the methods previously established are reviewed, assessed, and tailored to this specific task. The chapter aims to build upon these methods and develop a classifier in a previously unexplored field, being women's RU. The results of the classifiers are then discussed in addition to what this work means for the developing field.

Chapter 8: Optimising Reporting Locations for Females

The study reported here delves into the challenges posed by variations in sensor placement and reporting methods associated with IMG systems. These differences across systems have hindered direct comparisons of kinematic data and limit the adaptability of ML algorithms across different systems. To resolve these issues, suggestions have been made to standardise kinematic measurements, however, the adoption of these recommendations by manufacturers remains inconsistent. The research aimed to scrutinise the disparities in reported head acceleration and offer solutions to allow future research to provide accurate kinematic reports.

Chapter 9: Automated Epidemiology of Head Impact Events

This chapter explores two distinct classification tasks related to head impacts in women's collegiate rugby. The first task aimed to differentiate between HAEs with direct head contact (DHC) and those without direct head contact (NDC), while the second focused on predicting whether the impact occurred on the tackler or the BC. Each task utilised its unique dataset, with features extracted from various kinematic recordings. The process of developing models including feature selection, hyper-parameters optimisation and model selection is described. The performance of these classifiers is then reviewed with an analysis of how this work may affect the field outlined.

Chapter 10: Limitations, Conclusions, and Future Work

The thesis concludes with a review of the results and how they have answered the research questions formulated for this thesis. It provides context for the achievements of this work and how it will influence the field within future works. Obstructions to the work within this thesis are described as recommendations for improvements and future directions for this work.

2 Head Impact Telemetry in Sport

HIT systems are designed to measure the kinematic motion of an individual's head during collisions in sports and other high-risk activities (Arbogast et al., 2022). These systems are advancing the understanding of head impact kinematics which facilitates the extraction of valuable, actionable insights (Arbogast et al., 2022). Whilst these devices cannot diagnose brain injuries, the systems provide objective kinematic data regarding HAEs (Beckwith et al., 2018; Kieffer et al., 2020). This data can be utilised to identify potential injury risks in individuals and inform decisions about their well-being. HIT systems have been used in many sports to help coaches and medical staff monitor and manage the risk of head and brain injuries during collision events (Arbogast et al., 2022).

Within Chapter 1, there has been a brief introduction to IMGs, which are a type of head impact telemetry devices now commonly used in sports. This chapter provides greater information about HIT devices. In section 2.1, the history of the device is discussed from inception to the current field. Section 2.2 discusses how various HIT devices are used in modern-day sports. Section 2.3 discusses the reliability of HIT devices and why they were appropriate for World Rugby's recent mandate. The final section, 1.4, delves into the specific design considerations of IMGs that make them fit for purpose.

2.1 History of Head Impact Telemetry Devices

The concept of HIT dates back to the 1960's, in response to concerns about the prevalence of head injuries in American football (Patton, 2016; Schneider, 1961). The committee on the medical aspects of sports, created by the American Medical Association, suggested the collection of head acceleration data, to gain a greater understanding of the issue (Schneider, 1961). The first devices created were helmets containing tri-axial linear accelerometers, with one trialled within a professional American football game and the other during collegiate American football games over several seasons (Aagaard & Du Bois, 1962; Reid et al., 1971). An instrumented headband was also developed, designed to be worn and collect data from underneath a helmet (Moon et al., 1971). Data recorded with the instrumented helmet from collegiate games and with the headband showed peak linear accelerations over 1000 *g*, without the players suffering clear injury (Moon et al., 1971; Reid et al., 1971).

These values were considered to be more than what was believed to be safe limits at the time (Gurdjian et al., 1966). As the devices were improved and revised, the reported acceleration values began to decrease, although they were believed to still over-report the acceleration magnitudes. A revised instrumented helmet design reduced the magnitude of greatest linear acceleration events to around 400 *g*, whilst a study a decade after still reported values >500 *g* (Patton, 2016). Concurrently, T-shaped plates with accelerometer ensembles were adhered to the neck and worn within the mouth to measure the acceleration of the head in automotive and military studies (Patton, 2016).

It was not until the turn of the millennium that the HIT field began to gain significant interest, coinciding with the acknowledgement of the widespread nature of brain injury within the NFL (Omalu et al., 2005; Patton, 2016). Many studies were published within the first two decades of the millennium using helmets in American football, with papers reporting event frequencies, magnitudes, and even developing suspected concussion thresholds (Broglio et al., 2017; Le Flao et al., 2022; Zhang et al., 2004). Despite this initial promise, there were reports that helmets were not accurately recording head kinematics and that previous validation work had been unreliable (Jadischke et al., 2013). Despite this, many studies have continued to use helmet-mounted sensors (Le Flao et al., 2022). Alternatives to helmets were also developed, such as adhesive patches, headbands, and skullcaps, although these devices have proven similarly unreliable in testing (Kieffer et al., 2020).

With improvements in microtechnology, the first wireless IMGs were developed around 2010 (Paris et al., 2010; Patton, 2016) The design of IMGs means that device avoids the coupling issues that affect many of the previous devices mounted externally to the head (Knapik et al., 2007; Wright et al., 2021). Whilst they are not immune to recording issues they have consistently outperformed other devices both in the laboratory and on-field testing (Jones et al., 2022).

2.2 The Use of Head Impact Telemetry Systems in Sports

The use of the various modern HIT systems within the sporting field has been summarised in systematic reviews conducted by Patton et al., 2020 and Le Flao et al., 2022. Patton et al., (2020) analysed 168 studies that meet the inclusion criteria, all of which were published before December 31, 2019. The predominant sports investigated

were American football, accounting for 62% of the studies, followed by ice hockey (12%), soccer (11%), lacrosse (8.9%), RU/league (3.6%), Australian football (3%), and boxing (2%). The majority of research participants were collegiate level athletes (48%), with significant representation from high school (37%) and youth (22%) levels. Helmet-mounted sensors were the most prevalent (64%), followed by skin patches (24%), mouth guards (6.5%), headbands (4.2%), headgear (1.2%), caps (1.2%), ear-mounted sensors (0.6%), and mouthpiece-mounted sensors (0.6%).

Similar findings were reported by Le Flao et al., 2022, which included publications up until 31st December 2019. The majority focused on helmeted sports (75%), while the remaining studies covered non-helmeted sports (25%). The most studied sports were American football (57% of included studies), soccer (10%), and ice hockey (9%). The studied population was predominantly composed of collegiate (18-22 years) and high school (14-18 years) age groups, accounting for 77% of the total participants. Only 16% and 7% of the participants were from the youth (<14 years) and adult (>22 years) age groups, respectively. Additionally, there were 206 participants younger than 11 years of age reported in eight studies.

These findings revealed a significant male bias in the field, with 80% of the studies exclusively utilising male data. Among the remaining studies, 12% reported data for both males and females, while 8% focused solely on female participants. Of the 185 studies, 22% included one or more female participants, with a range of one to 58 female participants per study. Female participants represented less than 15% of the overall investigated population. They were best represented in soccer (366 participants), ice hockey (172 participants), and lacrosse (99 participants).

2.3 Measuring the Performance of Head Impact Telemetry Devices

There are two predominant issues with the reported data from HIT systems, which are:

i. Misestimating the kinematics of HAEs (Kieffer et al., 2020).
ii. Incorrectly recording spurious acceleration events (Kieffer et al., 2020).

The misestimating of kinematics may result from different sources. There may be issues with data processing methods, for example the translation to the head's CG, and alternatively it could arise from hardware issues. In order to measure a devices ability to accurately and reliably report data, they must undergo laboratory validation in

which their ability to correctly estimate kinematic is measured. The laboratory validation of HIT devices is discussed within 2.3.1

Spurious events can occur when recording data with IMGs when some vibration or impact not related to athletic incidents causes the sensor to accelerate and record an event. This is a noted problem, which has been reported as affecting all head impact telemetry devices. To address this, devices must be verified on-field, where their ability to correctly identify genuine HAEs is assessed against an observer is measured. This is discussed for various HIT devices in 2.3.2.

2.3.1 Lab Validation Results

Varying degrees of laboratory validation have been reported in previous studies, with some devices lacking proper validation altogether (Patton et al., 2020). Studies have also questioned the efficacy of these lab validation tests on some HIT devices (Jadischke et al., 2013). It was not until the consensus head acceleration measurements practice (CHAMP) guidance regarding laboratory validation was published that attempts to create a standard practice were created (Gabler et al., 2022). The recommendations outlined within the paper aim to create a standardised set of validation methods, to ensure are devices are reliable. The five key findings were:

- A wearable device that measures head acceleration must be independently validated for its intended application through controlled laboratory testing, and the laboratory should simulate the real-world loading environment in which the device will be used.
- Laboratory testing of wearable devices should use a validated biofidelic anthropomorphic test device (ATD) head form combined with a repeatable and reproducible test setup that enables testing across multiple levels of magnitude, duration, and direction that simulate on-field linear and angular head kinematics relevant to the setting of the study.
- Reference sensor setup and validation metric selection depend on the intended application of the wearable device, which can vary on four main levels: impact counting, impact magnitude, impact direction, and the time-history measurement of six-degree-of-freedom (6DOF) head kinematics.
- If a wearable device is designed to measure and report metrics derived from head kinematics, ground truth measurements must be collected with an ATD-

embedded laboratory-grade reference sensor system. If a wearable device is designed to count impacts only, a reduced reference setup enabling verification of impact events may be applied

- Processed data from the wearable device must be compared with ground truth measurements using validation metrics and statistical methods that enable complete, unbiased, and application-relevant assessment of accuracy and uncertainty.

The results from two studies that measured the performance of multiple HIT devices in laboratory settings will now be assessed. These studies were selected because of the range of devices included and the methodologies aligning with the criteria described by Gabler et al., 2022. Firstly, Kieffer et al., 2020 provides measurements of each of the most common HIT device types, whilst Jones et al., 2022 shows the testing used by World Rugby to inform their decision of which IMG to use in their recent mandate. Figure 2-1 depicts the experimental setup used within Jones et al., 2022.

Both studies measured reference kinematics using head forms with instrumentation packages consisting of three linear accelerometers (Endevco 7264b-2000; Meggitt Orange County, Irvine, CA) and a tri-axial angular rate sensor (DTS ARS3 Pro 18k; Diversified Technical Systems, Seal Beach, CA) at the heads estimated CG. Whilst these reference sensors have not been validated explicitly in the literature, the former is widely used in military, aviation and concussion research, whilst the later meets SAE J211, ISO 6487, NHTSA & FAA data acquisition requirements and the latest U.S. Government dynamic performance requirements(Diversified Technical Systems, 2024; Jones et al., 2022; Peregino & Bukowski, 2004; Walter, 2009). Both sensors record at frequencies of 20 kHz which is suitable for suitable for measuring impact mechanics.

Both papers use the Concordance Correlation Coefficient (CCC) for linear acceleration, rotational velocity, and/or rotational acceleration, allowing for easy comparisons to be made between studies. It quantifies concordance by incorporating measures of precision (closeness of repeated measures) and accuracy (closeness to the identity line), making it suitable for evaluating agreement between two continuous variables (Lin, 1989; Mahon, 2005). The specific reasoning for the selection of CCC by Kieffer et al., 2020 was that CCC may better account for bias (Lin, 1989). As

highlighted by Kusunoki et al., 2009, ICC may also be a suitable measure and can account for bias in certain circumstances. However, ICC and CCC are complementary rather than interchangeable, with CCC better suited for assessing agreement and ICC more focused on reliability (Lin, 1989; Shrout & Fleiss, 1979). This suggests more metrics could be used for measuring performance, but CCC is a suitable measure offering insight with a single measurement value.

A summary of test results from the two papers investigating the performance of various devices is shown in **Error! Reference source not found.**.

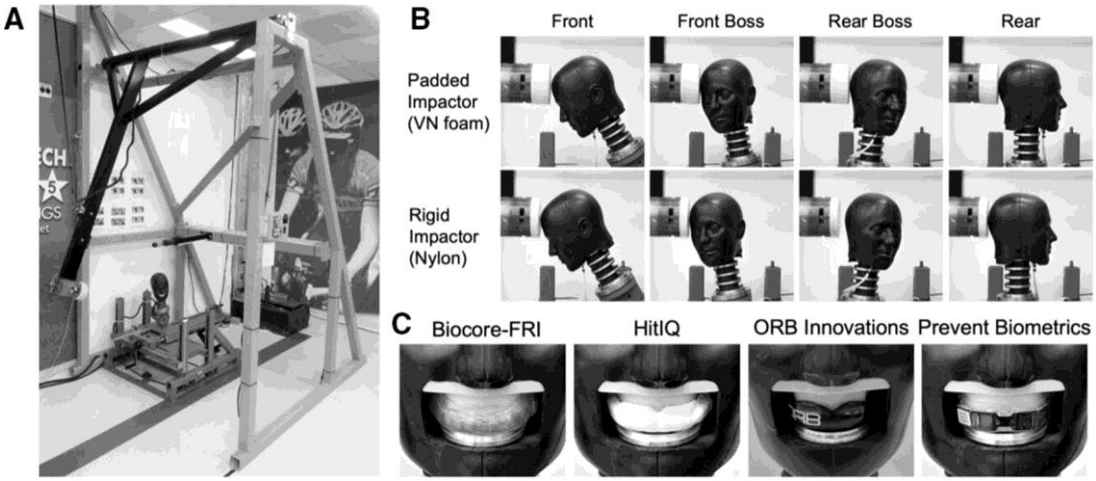
Table 2-1: Performance of various HIT devices in laboratory testing

Sensor Name	Sensor Type	Concordance Correlation Coefficient – (CCC)		
		Linear Acc.	Rotational Vel.	Rotational Acc.
xPatch	Patch	0.84(Kieffer et al., 2020)	1.00(Kieffer et al., 2020)	0.46(Kieffer et al., 2020)
SIM-G	Headband	0.48(Kieffer et al., 2020)	0.95(Kieffer et al., 2020)	0.39(Kieffer et al., 2020)
G-Force Tracker	Helmet	0.37(Kieffer et al., 2020)	0.32(Kieffer et al., 2020)	-
ORB Innovations	IMG	0.45(Jones et al., 2022)	0.53(Jones et al., 2022)	-
Prevent Boil-&-Bite	IMG	0.95(Kieffer et al., 2020)	-	0.97(Kieffer et al., 2020)
Prevent Custom	IMG	0.97(Kieffer et al., 2020), 0.98(Jones et al., 2022)	-	0.91(Kieffer et al., 2020), 0.98(Jones et al., 2022)
HitIQ	IMG	0.94(Kieffer et al., 2020), 0.94(Jones et al., 2022)	-	0.61(Kieffer et al., 2020), 0.98(Jones et al., 2022)
Biocore-FRI	IMG	0.98(Jones et al., 2022)	-	0.99(Jones et al., 2022)

Of the devices tested within these two studies (**Error! Reference source not found.**), IMGs recorded the highest measures for CCC in linear acceleration and rotational acceleration, excluding the ORB innovations device, which at the time was yet to be validated and withdrawn from the later rounds of testing (Jones et al., 2022). Of the non-IMG devices, the X-Patch performed best for LA, RV, and RA, although the measures of LA and RA are low when compared to the IMGs. The difference between RV and RA indicates that either there is noise in the RV signal that is magnified when the RA is calculated or that the methods of RA estimation are not suitable. As this device is adhered to a solid head form, this eliminates the skin artefact which has been shown to cause measurement errors with these devices. The SIM-G headband and G-Force tracker helmet recorded lower scores than the mouthguards in the three comparable scores, with the G-Force tracker helmet also recording a low CCC measure of RV. Similarly to the X-Patch, the SIM-G helmet showed much worse performance with RA than RV, indicating a propagation of errors or inaccurate estimation of RA. The poor results of these devices are potentially due to the coupling of the device to the head, explained in section 2.4.3.

Testing of the IMGs resulted in CCC measures greater than those recorded by other device types, indicating superior precision and accuracy. As the ORB IMG was not commercially available nor validated at the time of study it will be excluded from further analysis. In terms of linear acceleration, the devices performed consistently across the two tests, with the CCC values ranging from 0.94-0.98, with the highest scores indicating near-perfect agreement. For RA, CCC ranged from 0.61-0.99 with CCCs of four of the six tests greater or equal to 0.97. The lowest score was recorded by the HitIQ IMG in testing by Kieffer et al., 2020, which was before the device validation, which may indicate that data processing methods were changed between the two tests. Again a smaller difference is seen in the testing between the custom IMGs which may indicate differences between testing or device updates (Stitt et al., 2021). The remaining measures indicate near-perfect scores for CCC and the device's suitability for use in terms of validity. From this, we can see that in terms of laboratory validation, IMGs are accurate and precise, and the devices are suitable for use to accurately record kinematics in terms of laboratory validation.

Figure 2-1: Jones B, Tooby J, Weaving D, et al. Ready for impact? A validity and feasibility study of instrumented mouthguards. Br J Sports Med. 2022; 56(20):e0. doi:10.1136/bjsports-2022-105698. Reproduced from British Journal of Sports Medicine, Jones B, Tooby J, Weaving D, et al. (2022) with permission from BMJ Publishing Group Ltd. (A) Experimental set-up of pendulum impactor to simulate bareheaded impacts to the dummy headform for Phase I. (B) Padded (vinyl-nitrile) and rigid (nylon) impactor to the bareheaded dummy headform at the front, front boss, rear boss and rear locations of the headform. (B) The custom-fit instrumented mouthguard (iMG) mounted inside the headform with detachable three-dimensional printed detention.



Lab validation provides a "best case" scenario in terms of device validation. When worn by a user on-field numerous factors are introduced that can negatively affect the devices ability to record accurate data. To address this, on-field verification is discussed in the following section, while more specific IMG design considerations are discussed later in this chapter.

2.3.2 On-Field Verification Results

It is not possible to validate HIT devices on-field as it is not possible to establish ground truth kinematic measures with a reference sensor. This is an important consideration, as in-situ testing may cause noise that is accounted for in laboratory testing, the causes of, and solutions to, are discussed later in this chapter. Therefore, once devices are validated in a laboratory setting, they should undergo on-field verification to ensure that it is recording what it intends to record, HAEs, rather than spurious events. Spurious recording can be triggered in many ways, including, devices being removed, fitted, or displacing excessively due to poor fit (Jadischke et al., 2013; King et al., 2018; Wu et al., 2016). To provide a reference metric a ground truth is established by a reviewer who will record HAEs and compare their predictions to the outputs of the device.

Once data has been recorded, a metric must be used to establish the performance of the device against the ground truth predictions of the reviewer. Many appropriate metrics can be used to measure this performance, but for these tests, positive predictive value (PPV) and sensitivity are used, with more information available on these metrics in section **Error! Reference source not found.**. PPV reports the devices' ability to correctly report genuine events without including false positive events when compared to an observer identifying impacts. Sensitivity reports the relationship between the reports of true positive events and false negative recordings. Between these two metrics, it can be established whether all events reported are believed to be genuine, and whether the device has failed to record genuine events. The reality of on-field testing often reveals a trade-off between PPV and sensitivity, improving one metric typically results in a decline in the other. Relying on a single metric is insufficient. High sensitivity combined with low PPV leads to many spurious events being detected, while low sensitivity paired with high PPV results in many events being missed. More metrics could be used, but they are appropriate for the purpose.

Limitations of this method are the ground truth is only as good as the person who has verified impacts, which may influence scores. Whilst there is no better option currently, it can be mitigated by using an experienced rater. Computer vision may offer a solution to this problem in future, however, current models are not more reliably detecting HAEs than human reviewers (Mohan et al., 2024; Rezaei & Wu, 2022).

Table 2-2: On field performance of devices compares the performance of multiple HIT devices in on-field testing. The X-Patch patch-based sensor showed low PPV, with values reported as 16.3% (Kieffer et al., 2020) and 24.3% (Press & Rowson, 2017). Sensitivity was higher, with a value of 0.85 reported by Press & Rowson, 2017. Despite the high sensitivity, the low PPV indicates that the devices is likely triggered to record too easily, this would indicate even recording suspected to be genuine that a tendency would likely contain noise, and not be accurate measures. Despite laboratory validation, patch type sensors have been shown to include significant recording artefact. The helmet-mounted HITS device displayed relatively high PPV for active minutes, with an 88.0% value (Campbell et al., 2020). However, sensitivity was reported at 0.69 (Campbell et al., 2020), suggesting that while the system performed well in precision, it missed a significant number of true events. Another helmet-based device, the G-Force Tracker, achieved a lower PPV of 65.9% (Cortes et al., 2017).

Sensitivity data were unavailable, making it not possible to fully assess its reliability. The SIM-G headband sensor showed very low PPV at 6.1% (Campbell et al., 2020), indicating that it captured many irrelevant events. Similar to the X-Patch, high sensitivity was reported with the SIM-G headband (0.91) (Lamond et al., 2018). With the very low PPV, it would imply that the sensor is unreliable, and even recording concurrent with HAEs would be hard to trust as genuine given the high volume of spurious recordings.

Instrumented mouthguards (IMGs) showed a broad range of PPV and sensitivity scores in on-field testing. The device used within this thesis, the Prevent Boil-&-Bite exhibited a PPV of 55.0% for all minutes and 81.6% for active minutes (Kieffer et al., 2020). Sensitivity data were not reported, leaving some uncertainty regarding its ability to detect true events consistently. The custom fit Prevent IMG showed higher PPV values of 91.2% (Kieffer et al., 2020) and 89.0% (Jones et al., 2022) for all minutes. PPV increased to 96.4% and 94.0% (Kieffer et al., 2020; Jones et al., 2022) during active minutes. Sensitivity was reported at 0.75 (Jones et al., 2022), reflecting a good balance between precision and detection capability. Given the devices use identical sensor set ups, it is not unreasonable to expect the boil-and-bite version to perform comparably to the custom-fit version, assuming the device is well coupled to the wearer. The HitIQ device was reported to have a PPV of 60.0% for all minutes and 90.0% for active minutes (Jones et al., 2022). Sensitivity was relatively low, at 0.40 (Jones et al., 2022), indicating it might miss many true events despite reasonable precision. Lastly, the Biocore-FRI IMG demonstrated strong performance, with a PPV of 81.0% for all minutes and 98.0% for active minutes (Jones et al., 2022). Sensitivity was moderate, at 0.51 (Jones et al., 2022), again suggesting a better balance between detecting true events and avoiding spurious records could be struck.

Further analysis of the performance of all device types is included within the following sections.

Table 2-2: On field performance of devices

Sensor Name	Sensor Type	Positive Predictive Value (%)		Sensitivity
		Overall	Active Minutes Only	
xPatch	Patch	16.3(Kieffer et al., 2020), 24.3 (Cortes et al., 2017)	-	0.85 (Press & Rowson, 2017)
HITS system	Helmet	-	88.0 (Campbell et al., 2020)	0.69 (Campbell et al., 2020)
G-Force Tracker	Helmet	65.9 (Cortes et al., 2017)	-	
SIM-G	Headband	06.1 (Lamond et al., 2018)	-	0.91 (Lamond et al., 2018)
Prevent Boil-&-Bite	iMG	55.0 (Kieffer et al., 2020)	81.6 (Kieffer et al., 2020)	-
Prevent Custom Fit	iMG	91.2 (Kieffer et al., 2020), 89.0 (Jones et al., 2022)	96.4 (Kieffer et al., 2020), 94.0 (Jones et al., 2022)	0.75 (Jones et al., 2022)
HitIQ	iMG	60.0 (Jones et al., 2022)	90.0 (Jones et al., 2022)	0.40 (Jones et al., 2022)
Biocore-FRI	iMG	81.0 (Jones et al., 2022)	98.0 (Jones et al., 2022)	0.51 (Jones et al., 2022)

2.4 Instrumented Mouthguards

As the use of IMGs is a key theme of this thesis, the following sections will begin to more specifically focus on their use. To this point we have established they perform well in laboratory testing, but while one of the best performing devices in on-field testing, they still illustrate deficiencies. The following sections introduce key features of IMGs and how it influences their performance and use.

2.4.1 Mouthguards in Sports

Mouthguards are pieces of personal protective equipment designed to be worn on the teeth, to help protect the teeth and mouth from damage in sporting endeavours (Knapik et al., 2007). (Green, 2017) describes the three types of mouthguards manufactured, which differ in the fitting process used to personalise the mouthguard for the wearer's teeth. These three types are referred to as 'stock devices', 'boil-and-bite' or 'custom

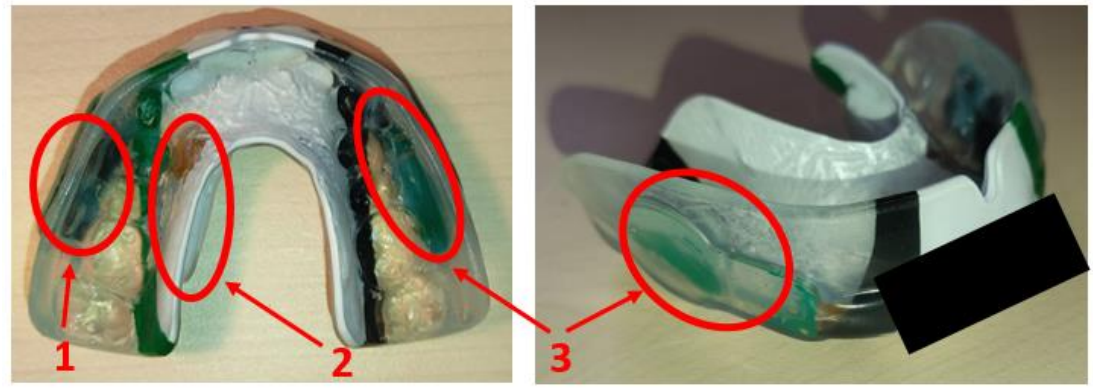
fit'. Custom devices are manufactured to fit a 3D mould of the wearer's teeth. Boil-and-bite type mouthguards are mass-produced from a standard template, before being heated and moulded to the wearer's teeth before use. Stock mouthguards require no moulding and are ready to use out of the box. Custom-fit mouthguards are typically the most expensive but offer the greatest protection from injury, followed by boil-and-bite devices then stock mouthguards. All devices offer increased injury protection in comparison to a person not wearing a mouthguard.

2.4.2 Fundamental Features of Instrumented Mouthguards

IMGs are mouthguards that have been enhanced with embedded inertial motion unit (IMU) sensors. They are capable of protecting the mouth from injury whilst also measuring head kinematic data. IMGs on the market are available as boil-and-bite or custom-fit options.

The IMU within an IMG typically consists of linear accelerometers and gyroscopes (Jones et al., 2022). These kinematic sensors will be embedded at different locations and orientations within the device, depending on the device manufacturer (Jones et al., 2022). Devices contain a mechanism for data storage and/or wireless transmission, allowing for recordings to be transmitted from the IMG to an end user's edge device in real time or after the event (Jones et al., 2022). These Data may be transmitted via radio frequency or Bluetooth depending on the manufacturer (Jones et al., 2022; Liu et al., 2020). The devices will also contain a battery and a wireless charging unit. To manage and regulate the device's functions, a microcontroller containing memory and processing units will make operational decisions and connect the peripheral units. An IMG is shown in **Error! Reference source not found.**, with key components highlighted.

Figure 2-2: Upper: An IMG with the motherboard and kinematic sensors (1), battery (2), and wireless charging unit (3) circled. Lower: an IMG being worn on the upper teeth and bony maxilla arch, with kinematic sensors visible.



2.4.3 Coupling of Instrumented Mouthguards to the User

To accurately represent the motion of the head, the HIT device and head must move in unison (Luke et al., 2024). Should the device be able to move freely about its position on the head, then the device is not truly capturing the motion of the head, but

only recording its motion about the head (Luke et al., 2024). The following text explains how this may be achieved from a mechanical perspective.

When coupled to a user's head, a HIT device may act as a damped harmonic oscillator system (Den Hartog, 1985). In this system, there are three main components. Firstly is a mass that oscillates harmonically when a force is applied. Secondly is a spring, which fixes the mass to a reference point and applies a force that opposes the motion of the device. Lastly is a damper, which disperses the kinetic energy of the mass as it moves, with Figure 2-4 depicting a simplified version of this system. As a force is applied to the mass to move it from its equilibrium point, the spring will return the mass to its equilibrium point whilst the damper reduces the size of future oscillations (Den Hartog, 1985).

The components of the system can be identified within a head-HIT device system (Wu et al., 2016). The mass can be considered to be the sensor housing and the method of fixing the sensor housing to the head as the spring. The damper is defined by the materials of the device and the method of fastening to the head. The head CG acts as a reference point from which the motion and relative motion can be described.

As an impact is recorded, the mass/sensor housing will move relative to the head, while the device's fastening mechanism will impede this motion. For example, in the case of a headband HIT device, the inertia of the sensor housing will be impeded by the elasticated headband material that holds the sensor to the head. A greater device mass will result in greater potential for high displacements, as are devices that are coupled with, or coupled to low stiffness materials (Den Hartog, 1985; Wu et al., 2016). By designing devices with low mass, high coupling stiffness, and high damping, such as IMGs, the chance of valid recordings increases as the potential for displacement decreases. To minimise the relative motion, devices should be designed with stiff coupling.

IMGs can achieve appropriate coupling to the head as they are worn on the upper teeth and the bony maxilla arch which offer little potential for displacement. This low level of displacement in turn leads to a more biofidelic measurement of motion, and greater accuracy compared to other in-situ recording methods (Wu et al., 2016; Greybe et al., 2020; King et al., 2016). This has been corroborated by studies researching the displacement of devices using high-speed cameras. An IMG was estimated to have <1

mm displacement from an ear canal reference point under loading, which was within video measurement error, and the lowest of the devices investigated (Wu et al., 2016). In comparison, patch and headband type HIT devices recorded displacements of two and four mm respectively (Wu et al., 2016). An image of this testing setup is shown in Figure 2-3: Test testing procedure for estimating sensor displacement during a head acceleration event. Reproduced with permission from Wu, L.C., et al. In Vivo Evaluation of Wearable Head Impact Sensors, Annals of Biomedical Engineering, 2015, Springer Nature.

Figure 2-3: Test testing procedure for estimating sensor displacement during a head acceleration event. Reproduced with permission from Wu, L.C., et al. In Vivo Evaluation of Wearable Head Impact Sensors, Annals of Biomedical Engineering, 2015, Springer Nature.

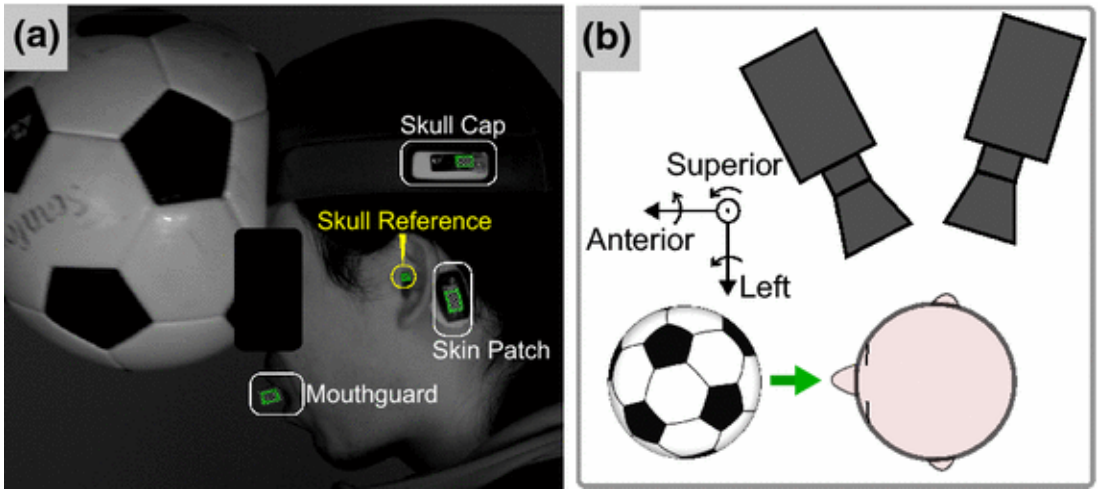
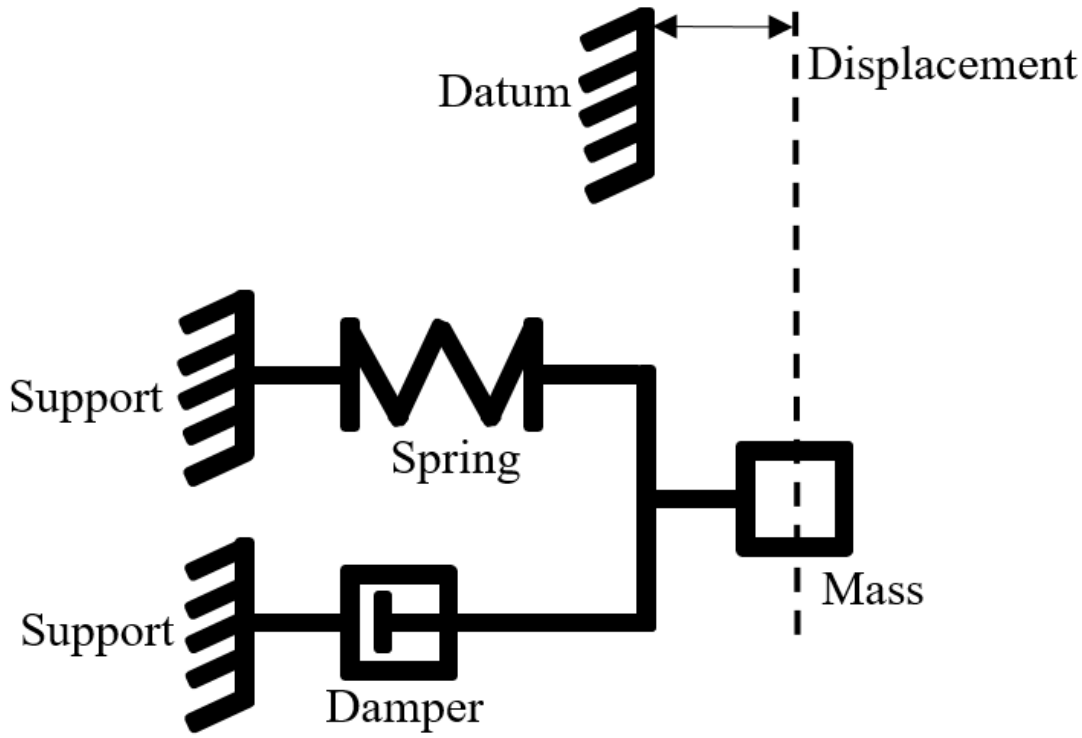


Figure 2-4: A single degree of freedom damped harmonic oscillator.



2.4.4 Sensor Placement

Another source of artefact is via the propagation of vibration though, with research suggesting the position of the sensors within the mouthguard may affect the accuracy of the results. Kuo et al. (2016) researched the effect of different mandible conditions on the accuracy of IMGs. In the testing, a Hybrid III anthropometric test device (ATD) was impacted with a jaw that was either removed, clenched, or free to move. In doing so it was seen that the free mandible condition could produce significant errors in the reported kinematics, with reported maximums of 40 and 80 % root mean squared error for rotational velocity and acceleration respectively. It was reported within the study that this could be reduced by relocating the sensor from the molar where the initial study was conducted to between the incisors, along with other minor design details, such as placing more material at the molar to alter the wearer's bite (Kuo, Wu, Zhao, et al., 2016). By making this change it has been reported that the data becomes of greater value to simulative study, reducing the error in predicted brain strain measurements versus brain strain calculated from ground truth anthropometric test device data. Currently, the locations of the sensors (accelerometer, gyroscope and

magnetometer), tend to vary between devices (Bartsch et al., 2014; Jones et al., 2022). For example, some devices have kinematic sensors placed over the molar, whereas other devices record data from the incisor (Jones et al., 2022). This will lead devices to be more sensitive to the effect of the mandible and therefore require different data processing methods in order to account for this.

2.4.5 Digital Signal Filtering

Raw data is the term given to the data recorded directly from a sensor before any processing or transformations take place. In the case of IMGs, for this data to be considered accurate or reliable, it must be digitally filtered to remove noise that distorts the signal. This is achieved using digital signal filtering methods and is used by most IMG manufacturers (Jones et al., 2022; Liu et al., 2020). Digital filtering converts the signal into the frequency domain where specific frequencies are removed, before the signal is converted back into the time domain.

The original IMG papers cite the use of the J211 crash test testing documentation to establish the filtering criteria for the first devices (Bartsch et al., 2014). Since the original paper's publication, many commercially available devices report the implemented low-pass Butterworth filters with cut-off frequencies in the range of 100 – 300 Hz (Jones et al., 2022; Liu et al., 2020). This means that frequencies below the value are retained or "pass" through the filter, whilst frequencies above are attenuated or removed.

Theoretically, this should remove spurious signals whilst retaining the kinematic information. This may not be the optimum approach, as mouthguards can carry signals from bites, shouts, and mandible interference which may not have been considered by the original documentation (Kuo, Wu, Hammoor, et al., 2016; E. M. P. Williams et al., 2021). To improve their approach, one company has reported the use ML algorithm to predict a suitable cut-off frequency for the Butterworth digital signal filter, by quantifying the noise within the recording and suggesting a filter of either 50, 100 or 200 Hz (Tooby et al., 2022a). An alternative suggestion is varying the threshold depending on the impact type and kinematic measure in order to improve the accuracy of the receding, although this move prove difficult in practicality (Gellner et al., 2024).

2.4.6 Proximity Sensors

Proximity sensors are infrared sensors that measure the strength of a reflected infrared signal from the wearer's upper dentition. As IMGs are discreet devices that can be easily worn and removed, monitoring when a player is wearing the device can be difficult. There is a concern that they might inadvertently record data when not on the user's teeth, potentially capturing non-relevant head accelerations. Proximity sensors can be used to monitor when they are and are not being worn which means they can stop recording when not worn. When no IR signal is recorded the device can stop recording, reducing the chance of spurious recordings.

The inclusion of proximity sensors aids in the prediction of partial or full device dislocation during impact (Luke et al., 2024; Wu et al., 2014). Changing proximity sensor readings are indicative of the device failing to remain coupled to the teeth, which will intern result in unreliable recordings as it is poorly coupled (2.4.3). This will help to reduce the prevalence of events such as biting, chewing and shouting as reported by Wu et al. 2014 and Williams et al. 2021 (E. M. P. Williams et al., 2021; Wu et al., 2014). This could be used to monitor devices fit over time as if mouthguards lose shape they may be prone to increase false positives or dislocations. Whilst this has not been proven, the higher performance of custom-fit mouthguards vs boil and bite may be indicative of this, as shown in **Error! Reference source not found.**.

In laboratory testing, proximity sensors have been shown to greatly reduce the number of falsely recorded events when compared to a linear acceleration threshold alone (Wu et al., 2014). In this setting a linear acceleration threshold of 10 g showed high sensitivity (92%) but low PPV (37%). When infra-red proximity was added to the linear acceleration threshold, PPV significantly improved to 92%.

2.4.7 Recording Trigger Thresholds

A recording trigger threshold is the value of a kinematic measure, above which, an IMG will record data. Should a HAE occur and the value is not exceeded, the trigger will not be met and the kinematic data will not be recorded. The purpose is to identify what could be events of interest for detecting mild traumatic brain injury or brain injury, whilst removing events that could be safe.

Linear acceleration is the measure most commonly used to set the trigger threshold. This partly due to the fact that there is an understanding of safe head linear

accelerations and that the alternative measure, rotational acceleration, is more computationally expensive to calculate. The values of linear acceleration that a human being can achieve through voluntary motion were estimated to be approximately 6 g for linear acceleration in laboratory testing (Ng et al., 2006). Additionally, g of 6, 8, and 10, have been recorded in roller coasters, non-contact RU events and trampolining without injury (Pfister et al., 2009; Sands et al., 2019; Tooby et al., 2024). By removing events of similar magnitudes then recordings relating to voluntary motion and safe intensity can be removed from the dataset.

Conversely, exposure to HAEs in the range of 10-15 g were predictive of CTE, depression, and other chronic health issues (Daneshvar et al., 2023; Montenigro et al., 2017). This illustrates that a transition appears to begin at 10 g from safe to more dangerous acceleration. This is reflected in the magnitude of trigger values commonly reported, which is generally between 5 and 13 g for, with a 10 g trigger being the most common (Tooby et al., 2024).

Whilst it aids the removal of non-interest events, Magnitude filtering alone is not a suitable method to filter genuine and spurious events. Studies have additionally reported collecting large numbers for false impacts when using threshold-only filtering. For example, in the study of Goodin et al. (2021), a head acceleration classification algorithm was developed, using data gathered from a HitIQ Nexus A9 (Goodin et al., 2021). Data was collected with only a magnitude filter and a reported 1,637 genuine events along with 12,075 spurious events were recorded (Goodin et al., 2021). If this method were to be used alone it would provide a PPV of 12 % should all values be assumed genuine events, which compares poorly with the reported values in **Error! Reference source not found.**, which considers mouthguards using an ensemble of methods to detect genuine events.

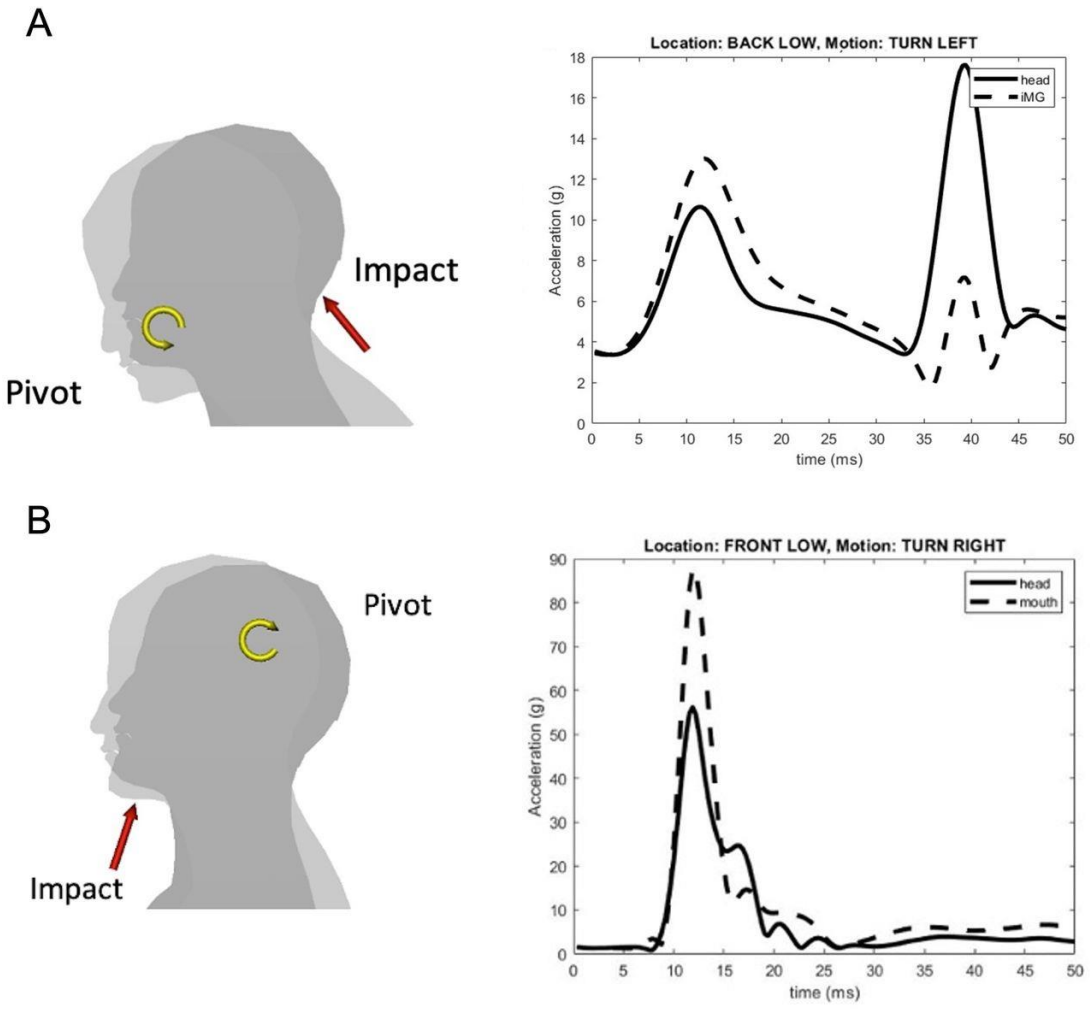
Whilst the use of a magnitude trigger is generally accepted, there is a growing body of research suggesting that it should be used carefully. Studies have shown that as linear acceleration vary across the head during a HAE, events with a linear acceleration magnitude of > 10 g at the head's CG may fail to trigger a recording (Wang et al., 2021).

During a HAE, the head acts as a rigid body, meaning rotational movement is constant across the body. Consequently, the linear acceleration across the head will vary,

dependant on the magnitude of rotational velocity and rotational acceleration. This means the head's CG may achieve an acceleration above the threshold, whilst the sensor remains at a level below this value and no recording is taken. This phenomena is illustrated in Figure 2-5.

This was dependent on the location of the impact to the head and the method of measuring acceleration (single axis vs magnitude). The result of this is a lowered specificity of HAE detection, which is an issue noted in table Table 2-2, regarding IMG HAE detection performance. It was reported that on occasions a 10 g linear acceleration at the CG was achieved in the head, but only 24.7-31.8% of these events will trigger a recording (T. Wang et al., 2021). Some impact locations resulted in a 30 g linear acceleration at the head's CG which still failed to trigger a recording (T. Wang et al., 2021). In order to mitigate this issue, an appropriate trigger values should be carefully considered. Whilst lower values will capture more genuine impact, more events not of interest will be captured. Whilst this is unlikely to present an issue with small sample sizes, as the dataset increases so will the required processing. Within this thesis, the methods used to account for this are discussed within Chapter 5.

Figure 2-5: Effect of the pivot point and direction of linear acceleration impulse may affect the PLA (A) Miss completely the impact peak or (B) Overestimate. iMG, instrumented mouthguard; PLA, peak linear acceleration. . Adapted from Influence of the frame of reference on head acceleration events recorded by instrumented mouthguards in community rugby players (Bussey et al., 2022). Reproduced with permission from BMJ Publishing Group Ltd.



2.4.8 Translating Linear Acceleration with Rigid Body Transformations

A problem in the HIT field is that different devices report linear acceleration from different locations, with some common locations reported within Table 2-3: Comparison of device reporting locations.. By reporting from a consistent location, it means that linear acceleration are not directly comparable between studies. Additionally, these difference will mean that ML algorithms trained on data from specific systems will not generalise to differing systems, due to the differences in data. In the CHAMP documentation, regarding reporting linear acceleration it is

recommended that kinematic measurements are rotated to match SAE J211 axes, linear acceleration is translated to the head's CG, and methods used to transform recorded data to analysable data are disclosed (Arbogast et al., 2022; Society of Automotive Engineers, 2003).

The kinematic data recorded from the sensor will be required to undergo rotation and translation in order to match the location and axes that the CHAMP documentation recommends reporting linear acceleration from (Bussey et al., 2022; Gabler et al., 2022). In order to perform the rotation and translation, precise measurements of the sensor recording axes relative to the SAE J211 axes are required, as are the distances between the recording location and the head's CG. The kinematic recordings may be rotated using analytical geometry, whilst a rigid body translation will allow for the acceleration to be calculated at the CG (Bartsch et al., 2014). The rigid body translation uses the following formula:

$$a_{CG} = a_S + \omega \times (\omega \times r_{CG}) + \alpha \times r \qquad (2\text{-}1)$$

In this equation the linear acceleration at the head's CG a_{CG} is calculated. This is calculated using the linear acceleration at the kinetic sensor a_S, the rotational velocity ω, the angular acceleration α, and the distance vector r from the kinematic sensor S to the CG.

Despite the best-practice being described by CHAMP as requiring the reporting locations and axes, companies still fail to offer the methods for practitioners to report the data in accordance with the guidance. In Chapter 8, the effect reporting from the sensor against an estimated wearers head CG is reported.

Table 2-3: Comparison of device reporting locations.

Device	Reporting Location	Reporting Axes
Prevent Biometrics (Bartsch et al., 2014)	Estimated 50th percentile head CG	SAE J211
SWA Protect (Y. Liu et al., 2020)	Molar	Sensor
Stanford (Y. Liu et al., 2020)	Incisor	SAE J211

2.4.9 Improving Instrumented Mouthguard Performance with Machine Learning

ML algorithms are a popular method of validating impacts in the HIT field, perhaps due to their convenience and performance (Patton, Huber, Jain, et al., 2020). In a review by Patton et al. 2020, 74% of the reviewed articles used some kind of algorithm to improve the prediction of genuine impacts.

Initially, questions were asked about the ability of the algorithms used to filter IMG data, with proprietary algorithms failing to reliably categorise the events as genuine or spurious (Patton, Huber, Jain, et al., 2020). This referenced the proprietary ML algorithms provided with devices from private companies (Patton, Huber, Jain, et al., 2020). The lack of performance was suggested to be due to the algorithms being developed from laboratory data which was not representative of on-field data (Patton, Huber, Jain, et al., 2020). This subsequently led to numerous studies reporting the development of algorithms specific to certain combinations of sports, sexes, and ages (Gabler et al., 2020; Wu et al., 2014). These algorithms were reported to perform better than proprietary algorithms, achieving near-human-level performance when verifying impacts (Raymond et al., 2022). To ensure that reliable data is available in all sports and settings, algorithms must continue to be made that can reply to classify data and expand the domain knowledge allowing future work to build upon it.

3 Computational Methodologies

3.1 Python and Object-Orientated Programming

The code developed for the studies within this thesis was developed using the Python programming language. Created in the 1980s, Python is an interpreted, interactive, and object-oriented programming language, which has gained popularity amongst programming languages for its power and high-level syntax (Sanner, 1999; Srinath, 2017).Python is used for a variety of tasks, including web development, scientific computing, data analysis, and artificial intelligence (Rossum, 2007; Srinath, 2017). Because of the popularity of Python and that it is open source, there are a vast number of libraries that offer pre-written code or increased functionality, allowing for faster and more powerful programmes (Srinath, 2017). Many of these libraries were used during code writing, to allow for the code to run more efficiently and quickly (Rossum, 2007; Sanner, 1999).

Python offers multiple data storage methods, each offering advantages depending on the requirements of a program or task (Rossum, 2007). Lists allow the storage of heterogeneous elements and are suitable for sequences of elements that might change in size or type, enabling easy manipulation using methods like appending, removing, or slicing (McKinney, 2011). Tuples are similar to lists, however, they are immutable, making them suitable for storing fixed collections or structures that should not change after creation (McKinney, 2010, 2017). Dictionaries use key-value pairs, offering fast access to data by associating unique keys with corresponding values and are valuable for organising and retrieving data efficiently (McKinney, 2017; Rossum, 2007). Arrays store homogeneous elements of a specific data type, optimising memory usage and providing efficient numerical computations (Harris, Millman, Walt, et al., 2020). The 'pandas' library offers a tabular data structure style data frames, capable of processing large volumes of structured data, with additional functionalities for data manipulation, cleaning, and analysis (McKinney, 2011, 2017).

Python functions are reusable blocks of code designed to accomplish specific tasks, applying a piece of logic without necessitating data storage (Oliphant, 2007). Their primary advantage lies in their reusability, allowing them to be executed multiple times and aiding in breaking down complex code into smaller, more manageable segments (Oliphant, 2007). By encapsulating specific functionalities into functions or

methods, code becomes more understandable, easier to debug, and simpler to update or enhance in the future (Lutz, 2013). Object-oriented programming (OOP) is a way of organising information and actions in a computer program by treating them as individual, self-contained entities called ‘objects’ (Phillips, 2010). The primary purpose of creating an object in OOP is to model and represent real-world entities, concepts, or elements within a software system (Rossum, 2000). OOP offers a way to organise code into reusable modular units, this allows for code to be easily re-used and maintained (Lutz, 2013). Each object includes data and methods whose purpose is to control the behaviour and actions of the object (Rossum & Drake, 2009). These behaviours or actions may include manipulation of the data, performing specific actions, or interacting with other objects.

There are four key principles within object-oriented programming: abstraction, encapsulation, inheritance, and polymorphism (Rossum & Drake, 2009). Abstraction is hiding the complicated inner workings of the object, allowing for objects to be viewed in terms of their behaviour, rather than the specifics of the code (Lutz, 2013). Encapsulation means wrapping the data and methods within an object, protecting them from outside interference (Phillips, 2010). Inheritance means that objects can inherit the properties and behaviours of parent classes, encouraging code reusability and the creation of specialised objects (Lutz, 2013). Polymorphism refers to the object's ability to take on many forms, allowing the object to behave differently in different contexts (Lutz, 2013; Rossum & Drake, 2009).

Examples of the functions used for coding are shown within the appendix, using the above methods. Functions allowed for slight changes to be made in the processing pipeline. For example, the data could be transformed at different frequencies with the change of a single variable rather than rewriting large blocks of code. Because of the interpretability of the data frame, they were the primary data storage and manipulation method used for the code, with example data shown within the appendix.

3.2 Action Recognition

Action recognition and human activity recognition are related fields in computer vision and artificial intelligence (Dang et al., 2020; Pareek & Thakkar, 2021). Both action and activity recognition involve identifying and classifying specific movement patterns performed by humans, often using video or kinematic recordings (Kong &

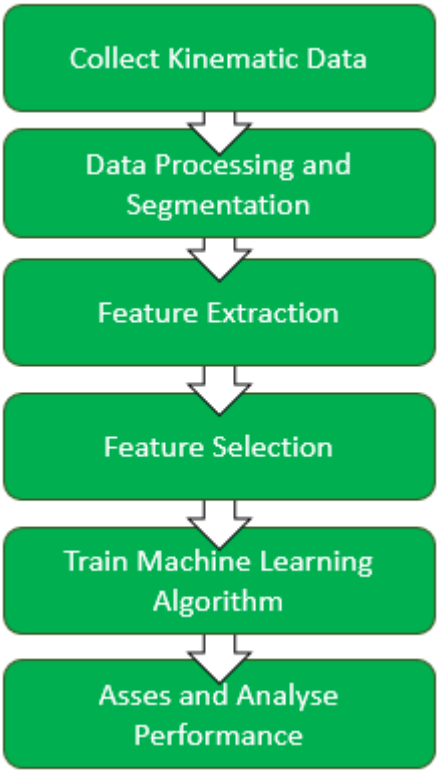
Fu, 2022). To conduct human activity recognition and action recognition, ML and deep learning techniques are commonly used (Dang et al., 2020; Herath et al., 2017; Yang et al., 2019). The choice of method used will depend on the complexity and nature of the activities being recognised and the amount of data available. Both areas have applications in various domains, including surveillance, robotics, healthcare, and sports analysis (Dang et al., 2020).

Action recognition focuses on the identification and classification of specific actions or gestures performed by individuals, for example, human walking, running, and sitting (Y. Liu et al., 2016). Action recognition is a valuable tool for sports analysts, having the potential to derive greater insights regarding facets of play than is possible manually. To date, it has been used in sports such as ballet, tennis, and volleyball to identify particular movement patterns (Hendry et al., 2020; Kautz et al., 2017; Tabrizi et al., 2020).

Activity recognition is a broad field that involves recognising and categorising various activities or behaviours performed by humans (Dang et al., 2020; Y. Liu et al., 2016). These activities are often more complex than individual actions and may include a sequence of actions or interactions between multiple individuals (Jobanputra et al., 2019). This may include recognising activities such as cooking, playing sports, and driving a car (Dang et al., 2020).

The studies reported in Chapters 7 and 9 focus on the identification and characterisation of specific patterns in kinematic data gathered during RU matches. These studies will use data processing pipelines and methodologies that are common in action recognition tasks. A slight difference is that action recognition typically requires the windowing of continuous kinematic data, IMGs only record very short, pre-segmented parts of the game during HAEs (Bartsch et al., 2014). The following sections will begin to explain in detail the methodologies used in and action recognition pipeline, with an example shown in Figure 3-1. The sections are ordered in to follow the steps of the pipeline.

Figure 3-1: An example of a human activity recognition pipeline



3.2.1 Defining Action Recognition

The process of action recognition can be defined using the following definition from Wang et al., 2019. Action recognition is formally defined in the following equation within the context of a user engaging in activities from a predefined set (A), where m denotes the total number of action types.

$$\mathrm{A} = \{\mathrm{A}\}_{\mathrm{i}=1}^{\mathrm{m}} \tag{3-1}$$

An activity sequence **s**, is captured through a series of sensor readings denoted as d_t at time t:

$$\mathbf{s} = \{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_\mathrm{t}, \ldots, \mathrm{d}_\mathrm{n}\} \tag{3-2}$$

The objective is to construct a model F that predicts activities $\widehat{\mathrm{A}}$ using the sensor data $\boldsymbol{s}$:

$$\widehat{\mathrm{A}} = \{\widehat{\mathrm{A}}_\mathrm{j}\}_{\mathrm{j}=1}^{\mathrm{n}} = \mathrm{F}(\mathbf{s}), \quad \widehat{\mathrm{A}}_\mathrm{j} \in \mathrm{A} \tag{3-3}$$

Whilst the ground truth activity is denoted as:

$$\mathrm{A}^* = \{\mathrm{A}_\mathrm{j}^*\}_{\mathrm{j}=1}^{\mathrm{n}}, \quad \mathrm{A}_\mathrm{j}^* \in \mathrm{A} \tag{3-4}$$

Where n is the length of the sequence and $n \geq m$.

The goal of human activity recognition is to create a model F that minimises the discrepancy between the predicted activity ($\widehat{A}$) and the ground truth activity (A^*). A typical performance metric used to evaluate the quality of the prediction is a loss function, $L(F(\mathbf{s}), A^*)$, aiming to minimise the resultant loss. The next section moves on to discuss the methods used within Chapter 6 in order to generate and select features. By providing features that allow the machine learning algorithm to discriminate between the different groups, the algorithm is capable of minimising the resultant prediction loss.

3.3 Data Processing, Feature Development, and feature Selection

An adage in ML is "Garbage in, garbage out", the essence of which being that if algorithms are trained with badly prepared data, the models with return bad results (Kilkenny & Robinson, 2018; Vidgen & Derczynski, 2020). In order to ensure the classifiers developed in Chapter 7 and 9 are reliable, this sections reports the features used to create the training and testing datasets, as well as the methods used to select features and select generalisable features.

3.3.1 Features Development

As the domain of genuine impact detection in HIT is well developed, the novel studies in this thesis were capable of building upon this existing body of literature. The feature groups in the following have been develop either in line with the methodologies of previous studies, or inspired from the feature groups used (Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2014).

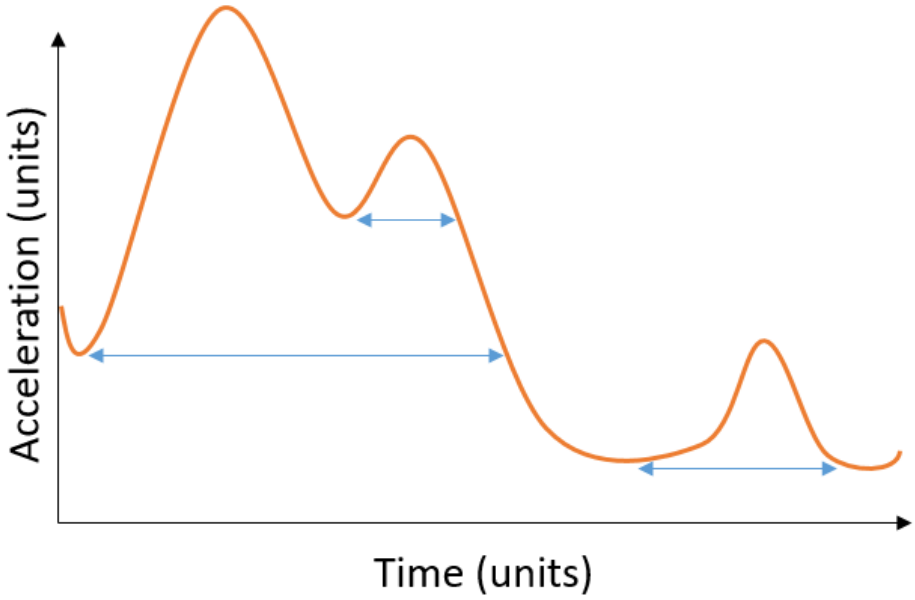
Pulse Parameters

The analysis aimed to identify pulses in the kinematics signal, considering them as regions bounded by contour lines. The objective of this feature group was to capture transient frequencies, providing insights into the intensity and duration of the primary impact peaks.

The process began by identifying a value, V, in a 1-dimensional array where the surrounding data points had lower values than V, defining it as a maximal value. To gauge the full size of the pulses, the count of data points with declining values was measured. If the values didn't decline, the pulse was considered ended. The minimum

number of data points before the gradient ceased to decline was regarded as half the pulse width. At this position, the lowest bounding contour line was established, and from this line, the peak prominence was measured. In cases of multiple peaks within a signal, the maximum values for both prominence and width were considered. Criteria may be set to use only peaks of a certain width or prominence. This is illustrated in **Error! Reference source not found.**. The identification of pulse characteristics was designed to provide a group of measures that could capture information about the signal pertaining to both continuous and discrete frequencies contained within, along with pseudo measurements of their power. This could be interpreted from the frequency, width, and prominence of the pulses.

Figure 3-2: An illustration of how pulses may be detected within a kinematic signal.



Power Spectral Density

Power Spectral Density (PSD) is a measure that describes how the power of a signal is distributed across its frequency components (Solomon Jr, 1991). It is a representation of the power per unit frequency, indicating the strength of a signal at different frequencies. PSD is commonly used in fields like signal processing and communication systems to analyse the frequency content and characteristics of a signal. PSD based features have been reported in most HIT impact detection algorithms, providing value features in terms of classifying head impacts (Goodin et

al., 2021; Wu et al., 2014). They have also been reported to be used in many other action and activity recognition studies (Dang et al., 2020).

Various methods exist for calculating PSD, but for this thesis, Welch's method was employed. Welch's method is an approach for estimating the PSD of a time series signal, providing insights into the underlying continuous frequencies within the signal. This technique relies on periodogram spectrum estimates, often implemented through fast Fourier transformations, facilitating the conversion of signals from the time domain to the frequency domain.

What sets Welch's method apart is its segmentation approach. The data is partitioned into overlapping segments, which are subsequently windowed to mitigate spectral leakage, smoothing the data at the signal edges. Following segment creation, a periodogram is computed using fast Fourier transformations on each segment. The average PSD is then determined across all signals, serving as an estimate of the frequency components. Consider a signal represented by $x[0], x[1], \ldots, x[N-1]$, and the segments as follows:

Segment 1: $x[0], x[1], \ldots, x[M-1]$

Segment 2: $x[S], x[S+1], \ldots, x[S+M-1]$

...

Segment K: $x[N-M], x[N-M+1], \ldots, x[N-1]$

Where M represents the number of points in each segment (or batch) of data, S is the number of points to shift between segments, and K is the total number of segments or batches. For each segment, the data is windowed, and a discrete Fourier transformation is calculated for some frequency when $v = i/M$ with $-(\frac{M}{2}-1) \leq i \leq \frac{M}{2}$:

$$X_k(v) = \sum_m \mathrm{x[m]w[w]}e^{-j2\pi vm} \tag{3-5}$$

Where:

$$m = (k-1)S, \ldots, M + (k-1)S - 1 \tag{3-6}$$

And $w[m]$ is the window function. For each segment (k=1 to K), the modified periodogram value $P_k(f)$ is calculated from the discrete Fourier transform:

$$P_k(v) = \frac{1}{W}|X_k(v)|^2 \tag{3-7}$$

Where:

$$W = \sum_{m=0}^{M} \mathrm{w}^2[m] \tag{3-8}$$

And $P_k(v)$ is the modified periodogram for the kth segment. The average of the periodogram values is then obtained to provide the Welch's estimate of the PSD.

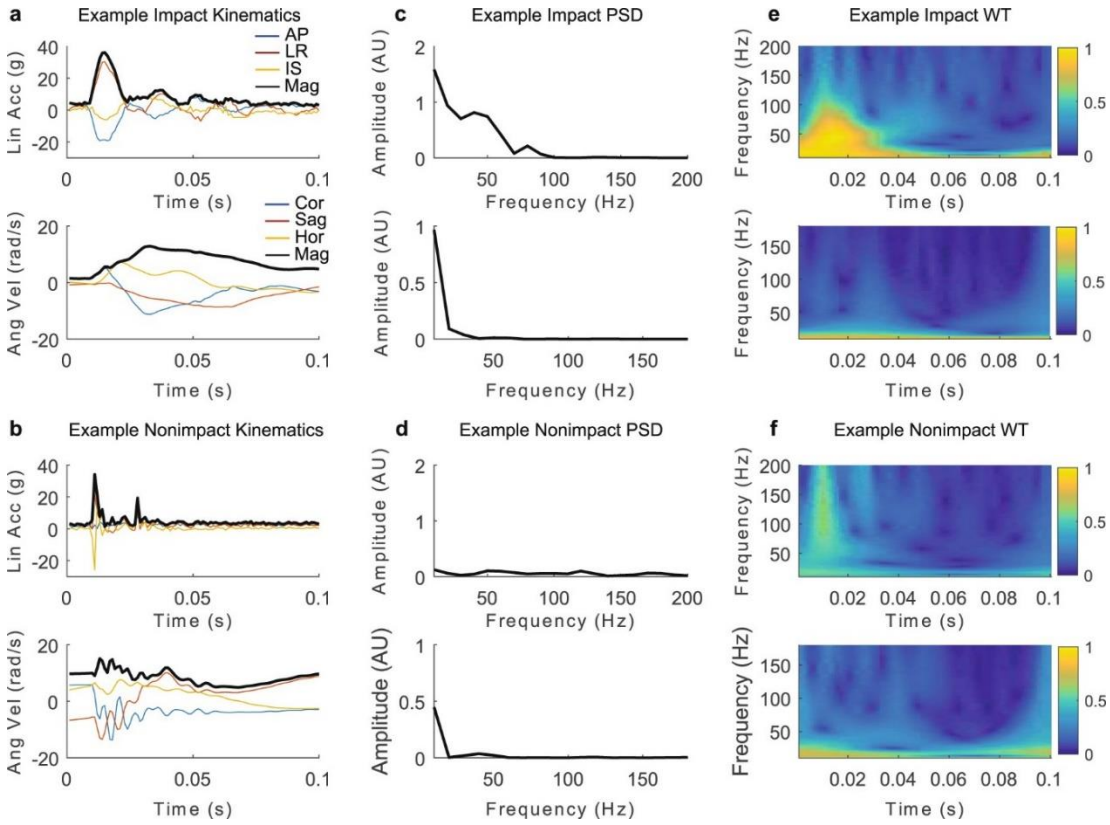
$$S_x(v) = \frac{1}{K\sum_{k=1}^{K} P_k v} \tag{3-9}$$

Increasing the number of segments and averaging results enhances the consistency of the PSD estimate during analysis. While short signals may benefit less due to fewer creatable segments, increasing the number of overlapping data points compensates for this limitation. The number of segments can be increased by augmenting the number of overlapping data points. Although this heightens the prevalence of redundant information, the application of non-linear windows helps mitigate the impact, ensuring the reliability of the analysis.

Wavelet Transformations

Continuous wavelet transformation (CWT) is a mathematical tool used in signal processing to analyse the frequency content of a signal over time (Rioul & Duhamel, 1992; Ryan, 1994; Yamada, 2006; D. Zhang, 2019). Unlike the discrete Fourier transform (DFT), which provides frequency information across the entire signal, the CWT aims to identify localised frequency information in both time and frequency domains (D. Zhang, 2019). Employed widely across various fields, including signal processing, image processing, and feature extraction, the CWT provides insights into localised frequency characteristics within a signal (Yamada, 2006). Wavelet transformation have been used for the purposes of action recognition and HAE detection, with low frequency wavelets (10-30 Hz) proving particularly powerful in the classification of the later (Dang et al., 2020; Wu et al., 2018). Examples of how CWT and PSD analysis would interpret spurious and genuine HAE recordings are shown in Figure 3-3.

Figure 3-3: Kinematics, PSD, and WT of an Example Impact and Nonimpact. Example impact (a) and nonimpact (b) kinematics show qualitative differences between these two events, with the nonimpact exhibiting higher frequency impulses and oscillations. Such frequency-domain differences are reflected in the Fourier transform power spectral density (PSD) plots (c and d) and wavelet transform (WT) plots (e and f), where color represents amplitude. Figure from: Wu, L.C., Kuo, C., Loza, J., et al. (2018). Detection of American Football Head Impacts Using Biomechanical Features and Support Vector Machine Classification. Scientific Reports, 8, 855, Figure 4. https://doi.org/10.1038/s41598-017-17864-3. Licensed under CC BY 4.0.



A mathematical explanation is offered by (Cohen & Kovacevic, 1996). The CWT operates by decomposing a continuous-time signal into its constituent frequency components. It measures the similarity of the input signal to a scaled and translated version of a fundamental waveform called the "mother" wavelet function, represented as ψ(a, b). The mother wavelet is a basic waveform that serves as the building block for the CWT. It determines the characteristics of the wavelet transform and defines the shape and behaviour of the wavelets used in the analysis. This function, characterised by continuous parameters of scale 'a' and position 'b', undergoes correlation with the input signal across different scales and positions, thereby representing the signal's frequency content in both time and frequency.

The scale 'a' parameter in the CWT corresponds to the dilation or compression of the mother wavelet. Larger scale values represent broader wavelets, capable of capturing

lower-frequency information, while smaller scale values generate narrower wavelets, suitable for detecting higher-frequency details. The position 'b' parameter controls the translation or shift of the wavelet function along the signal's time axis, enabling the analysis of different temporal regions.

In essence, the CWT examines how well the signal matches scaled and shifted versions of the mother wavelet across various scales and positions, enabling the identification of localised frequency characteristics within the signal. Employed widely across diverse fields such as signal processing, image analysis, and pattern recognition, the CWT provides valuable insights into the localised frequency features present in complex signals, enabling enhanced understanding and analysis of dynamic phenomena.

For a continuous signal 'f(x)' and the mother wavelet function ψ (a, b), the CWT formula is expressed as[2]:

$$T_{\psi}(a,b) = \frac{1}{\sqrt{C_{\psi}}} \int_{-\infty}^{\infty} \psi^{(a,b)}(x) f(x) dx \tag{3-10}$$

Here, C_{ψ} is a normalisation constant ensuring the admissibility condition is met, and the integral represents the continuous convolution of the signal $f(x)$ with the scaled and translated wavelet function ψ (a, b).

The Ricker wavelet, also known as the Mexican hat wavelet, is a specific wavelet used in CWT. Initially introduced in seismic signal analysis in 1984, it has since found extensive applications in diverse research domains, as defined below (Ryan, 1994):

$$Ricker = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2} \tag{3-11}$$

The CWT facilitates signal decomposition into time, scale, and potential direction, employing localised kernel functions with continuous dilation and translation parameters. This transformation can be likened to a windowed Fourier analysis, allowing window changes in length and position within the original signal. Consequently, wavelets offer valuable local frequency information from recorded events. The one-dimensional wavelet function $\boldsymbol{\psi}^{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x})$ are defined as:

$$\boldsymbol{\psi}^{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{x}) = \frac{1}{\sqrt{|a|}} \boldsymbol{\psi}\left(\frac{\boldsymbol{x} - \boldsymbol{b}}{\boldsymbol{a}}\right) \tag{3-12}$$

For a CWT, where the parameters (a, b) are continuous, the analysing wavelet $\boldsymbol{\psi(x)L^2}$ $(\boldsymbol{R})$ must satisfy the following admissibility condition:

$$C_{\boldsymbol{\psi}} \equiv \frac{\int_{-\infty}^{\infty} \left|\widehat{\boldsymbol{\psi}}(\boldsymbol{\omega})\right|^2}{|\boldsymbol{\omega}|} d\boldsymbol{\omega} < \infty \tag{3-13}$$

Where $\widehat{\boldsymbol{\psi}}(\boldsymbol{\omega})$ signifies the Fourier transform of ψ(x):

$$\widehat{\boldsymbol{\psi}}(\boldsymbol{\omega}) = \int_{-\infty}^{\infty} e^{-i\omega x}\, \boldsymbol{\psi}(\boldsymbol{x}) \boldsymbol{dx} \tag{3-14}$$

The admissibility condition implies that the wavelet function $\boldsymbol{\psi}(\boldsymbol{x})$ must lack a zero-frequency component, specifically $\widehat{\boldsymbol{\psi}}(\mathbf{0}) = \mathbf{0}$, under some mild conditions on its decay rate at infinity. The CWTs and the inverse transformation of the data function $\boldsymbol{f}(\boldsymbol{x})$ belonging to $\boldsymbol{L^2}(\boldsymbol{R})$ can be mathematically defined using the CWT formula, and the formula as follows:

$$f(x) = \frac{1}{\sqrt{C_\psi}} \int_{-\infty}^{\infty} T_\psi(a,b) \psi^{(a,b)}(x) \frac{da\, db}{a^2} \tag{3-15}$$

3.3.2 Feature Selection: Maximum Relevance, Minimum Redundancy

Feature selection plays an important role in the construction of ML algorithms. By using a feature reduction method to remove non-valuable features, model performance can be increased, training times can be reduced, and interpretability can be improved (García et al., 2015). The minimum redundancy maximum relevance (mRMR) method has been developed as a solution to the problem of developing an optimal feature set, which is considered to be an np-complete problem (Zhao et al., 2019a). This is done by calculating the predictive power of the feature, whilst controlling for redundancy within the group of predictive features (Zhao et al., 2019b).

Zhao et al., 2019a defines the mRMR method as the following. Assuming a total of m features, the importance of a given feature $X_i (i \in \{1, 2, \dots, m\})$ can be expressed as (Zhao et al., 2019a):

$$f^{mRMR}(X_i) = I(Y, X_i) - \frac{1}{|S|} \sum\nolimits_{X_{s \in S}} I(X_s, X_i) \tag{3-16}$$

Where Y is the class label and S is the selected features, whilst |S| is the number of features in the original feature set. Additionally, $Xs \in S$ represents one feature from

the set S, whilst X_i denotes a feature not currently selected. The function *I(,.,)* is the mutual information:

$$I(Y,X) = \int_{\Omega_Y} \int_{\Omega_X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) dx\, dy \tag{3-17}$$

Where **Ω Ω** are representative of the sample spaces of Y and X, p(x, y) is the joint probability density and p() is he marginal density function. The mutual information formula changes for discrete variables Y and X to:

$$I(Y,X) = \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3-18}$$

At each stage of the mRMR process, the feature that maximises the feature importance score $max_{X_i \notin S} f^{mRMR}(X_i)$ will be added to the feature section set S.

This is adapted however for the frequency correlation quotient (FCQ) variant of the mRMR method, where the method changes to:

$$f^{FCQ}(X_i) = \frac{F(Y,X_i)}{\frac{1}{|S|}\sum_{X_s \in S} \rho(X_s X_i)} \tag{3-19}$$

Here, the term $F(Y,X_i)$ is the F-statistic and $\rho(X_s X_i)$ is representative of Pearson's correlation coefficient, giving the method the name the F-Statistic Correlation Coefficient.

The FCQ variant has been previously empirically tested and shown good results in training all forms of classification models. For some classifiers, such as random forest, other feature selection methods can produce greater accuracy in classifiers, although these are far more computationally expensive. This can be negated in some scenarios and presenting more features to the random forest can improve performance.

3.3.3 Outlier Isolation

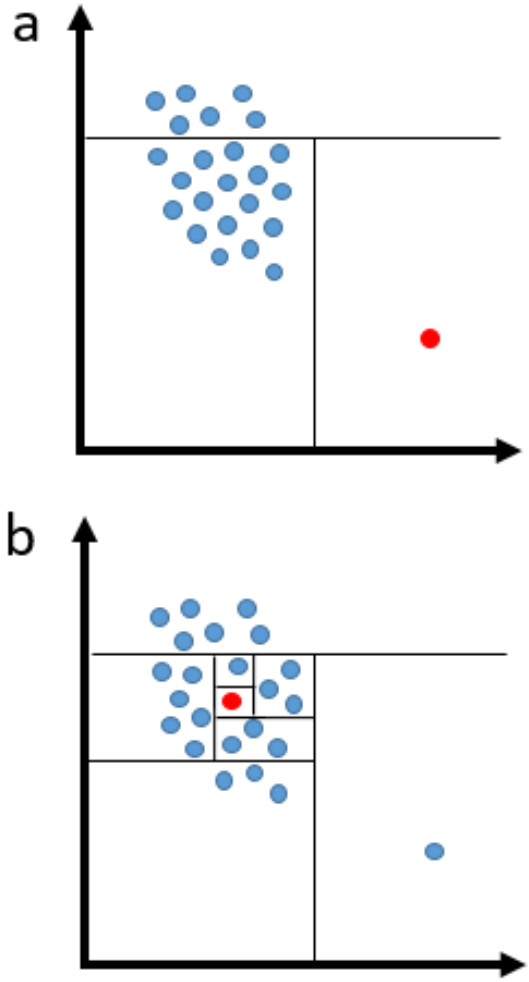
Anomaly detection involves identifying data points that significantly deviate from expected norms or patterns within a dataset (Buschjäger et al., 2022; Liu et al., 2008a). It is commonly applied in fields such as cybersecurity, healthcare, manufacturing, and environmental monitoring (Buschjäger et al., 2022). In this thesis, the focus was on leveraging anomaly detection techniques to identify outliers in smaller datasets. This

process aims to refine the dataset, ensuring its alignment with expected norms, thereby enhancing its generalisability to new data.

Isolation forests were developed as a robust method designed to effectively isolate anomalous instances. Unlike previous methods that relied on establishing norms before flagging outliers, isolation forests pioneered a method that directly identifies anomalous data, thereby reducing computational complexity (F. T. Liu et al., 2012). The essence of isolation lies in separating instances, while anomalies, being few and distinctive, are more susceptible to isolation.

Conceptually akin to random forests, an isolation forest comprises multiple decision trees (Liu et al., 2008a). Each tree's goal is to randomly segment the data until a single instance of the data is uniquely isolated at a leaf node, as shown in Figure 3-4. (Liu et al., 2008a). The core idea behind isolation forests is that data fed into these trees necessitates fewer partitions to reach a leaf node (Liu et al., 2008a). Given a specific feature, partition values are chosen randomly between the maximum and minimum feature values (Liu et al., 2008a). The tree is terminated when one of the following criteria is met, either the tree reaches a predefined path length limit, each point is isolated into its own leaf, or the data contained within each leaf consist of the same value (Liu et al., 2008a). The total path length, defined as the number of decision nodes traversed from the root to the leaf node, serves as a pivotal metric within the isolation forest (Liu et al., 2008a).

Figure 3-4: An illustration of how many repetitions it may take an isolation forest to separate an outlying data point (a) vs. normal data point (b).



More formally defined, an isolation tree works in the following way. With the data sample $X = \{x_1, x_2, \ldots, x_n\}$ of length n in a multi variate normal distribution, an isolation forest recursively splits the sample X, with a random test value of p from within the limits of the attribute q (Liu et al., 2008a). If T is the node of an isolation tree, T can either be in internal node, with a test and exactly two children (T_l, T_r), or an external node with no children nor test (Liu et al., 2008a). With the assumption that all data points are distinct, each data point is isolated at a separate leaf node when the tree is fully grown (Liu et al., 2008a). This will incur a tree where the number of leaves is equal to n and the number of decisions is equal to $n - 1$ (Liu et al., 2008a).

The path length, denoted as $h(x)$, represents the number of edges traversed from root to leaf nodes. The abnormality score ($s(x,n)$) is computed based on the average path length across a collection of isolation trees. It is calculated using the formula:

$$s(x,n) = 2^{-\left(\frac{E(h(x))}{c(n)}\right)} \quad (3\text{-}20)$$

$E(h(x))$ is the average height of a particular instance x across multiple isolation trees (Liu et al., 2008a). This average height indicates how 'isolated' or 'deep' the instance is within these trees (Liu et al., 2008a). Anomalies are expected to have shorter average path lengths (Liu et al., 2008a). Additionally $c(n)$ represents the average height of terminal nodes in the isolation trees for a dataset of size n (Liu et al., 2008a). It serves as a normalization factor to scale the height of x in relation to the general behavior of the trees (Liu et al., 2008a).

The anomaly score ranges between 0 and 1, where values close to 1 indicate near-certain anomalies, values smaller than 0.5 suggest typical instances, and approximately 0.5 implies the absence of distinct anomalies across the entire sample (Liu et al., 2008a).

3.4 Machine Learning Algorithms

An introduction to ML can be found in section 1.6, which offers a description of the most fundamental concepts of ML. Within this section, different sub-groups of ML algorithms will be introduced. The underlying methods of the algorithms will be discussed, and how they may be used to add value to the thesis.

3.4.1 Classical Machine Learning and Deep Learning Algorithms

ML has emerged as a crucial tool for human activity recognition, offering high accuracy and efficiency in detecting patterns in data (Jobanputra et al., 2019; Kong & Fu, 2022; Lara & Labrador, 2013). This section discusses the use of classical ML and deep learning and how their characteristics influence their suitability in the thesis.

Deep learning algorithms, a subset of ML, are characterised by their use of neural networks with multiple layers that enable hierarchical learning of data representations (Goodfellow et al., 2016; LeCun et al., 2015). These algorithms excel at modelling complex, non-linear relationships by progressively extracting abstract and intricate features as data passes through successive layers, a process known as learning "deep

features" (X. W. Chen & Lin, 2014). Unlike traditional ML methods that require explicit feature engineering, deep learning algorithms can be trained directly on raw time-series data, automating feature extraction and simplifying data preparation (Janiesch et al., 2021). Moreover, they inherently filter noise, retaining only meaningful features, which enhances the model's ability to identify underlying patterns (Janiesch et al., 2021).

Despite these positives, deep learning algorithms have limitations. These models are computationally intensive and require large datasets to effectively learn intricate patterns without overfitting (LeCun et al., 2015). With smaller datasets, overfitting is a common challenge, as the model may memorise training data rather than learn generalisable patterns that apply to unseen examples (Salman & Liu, 2019). Techniques such as data augmentation, transfer learning, and regularisation can mitigate these issues by increasing the effective dataset size, while transfer learning applies knowledge from other tasks to a new challenge (Su et al., 2024; Um et al., 2017). Nonetheless, deep learning models often remain less interpretable than traditional methods, as their learned features are abstract and not easily mapped to domain-specific knowledge (Rudin, 2019). One form of deep learning is perhaps more appropriate for use with small datasets however, the multilayer perceptron (MLP) generally requires less data to train effectively compared to other networks as MLPs are simpler in architecture. MLPs are discussed further in 3.4.3.

Classical ML algorithms typically require data to be transformed into a vector of descriptive features for the algorithm to interpret (Goodfellow et al., 2016). This requires domain knowledge to create valuable features to allow the model to determine important patterns (Shalev-Shwartz & Ben-David, 2013). This process can be time-consuming and labour-intensive should these features not be well defined within the domain, when compared to deep learning algorithms, capable of generating features (Alpaydın, 2013; Dang et al., 2020). Other benefits of using classical ML methods include that less data is required to create reproducible results, reducing the time and cost of data collection (Dang et al., 2020). The use of classical ML models generally leads to more interpretable results, with non-abstract features allowing for a greater understanding of the mechanisms that lead to the model’s predictions (Rudin, 2019). Further methodologies have been developed to help interpret these and more complex

models in a movement known as explainable AI (Lundberg & Lee, 2017; Minh et al., 2023).

The following methods are described in the following text, with the chapters that are relevant to following the term.

- Decision tree,
- Multi-Layer Perceptron
- Support Vector Machine
- Logistic Regression

3.4.2 Decision Tree

Decision trees are a popular and versatile machine-learning algorithm used for both classification and regression tasks (Alpaydın, 2013). The decision tree algorithm is trained by making decisions by recursively splitting the dataset into subsets based on the values of input features (Rokach & Maimon, 2005). The goal is to create a tree-like structure where each internal node represents a decision based on a specific feature, and each leaf node represents the predicted output (Tangirala, 2020). A decision tree begins at the root node, which includes the entire dataset (Kingsford & Salzberg, 2008). The algorithm evaluates different features to determine the one that best splits the dataset into subsets that are more homogenous in terms of the target variable. Common measures for evaluating the best split include Gini impurity, entropy, or information gain (Alpaydın, 2013). Tangirala, 2020 defines the calculation of information gained from a partition p as:

$$IG(Y_{train}, p) = I(Y_{train}) - I(Y_{train}|p) \tag{3-21}$$

Where $IG(Y_{train}, c)$ represents the information gained when the dataset Y_{train} is partitioned using the partition p. Additionally, $I(Y_{train}|p)$ represents the information contained within the data after being split by criteria p, whilst $I(Y_{train})$ is the information prior to the split.

The Gini index and entropy, also assess data purity by gauging the evenness of class distribution (Tangirala, 2020). A classifier aims to minimise these scores by reducing randomness and thereby increasing purity. Formally, Gini and entropy are defined in the following (Tangirala, 2020):

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \quad (3\text{-}22)$$

$$Entropy = \sum_{i=1}^{n} -p_i log_2(p_i) \quad (3\text{-}23)$$

Here, n is the number of classes, and p_i is the probability that a randomly selected data point from label i will be correctly assigned to its label (Tangirala, 2020).

The dataset is divided into subsets based on the chosen feature (Kingsford & Salzberg, 2008). Each subset corresponds to a branch from the root node to a child node. The process is repeated recursively for each subset, treating them as separate datasets (Rokach & Maimon, 2005). The algorithm selects the best feature for each subset and splits it further. For categorical features, the tree considers all possible values, whilst for numerical features, the algorithm selects the optimal threshold that maximises purity when splitting the data (Tangirala, 2020).

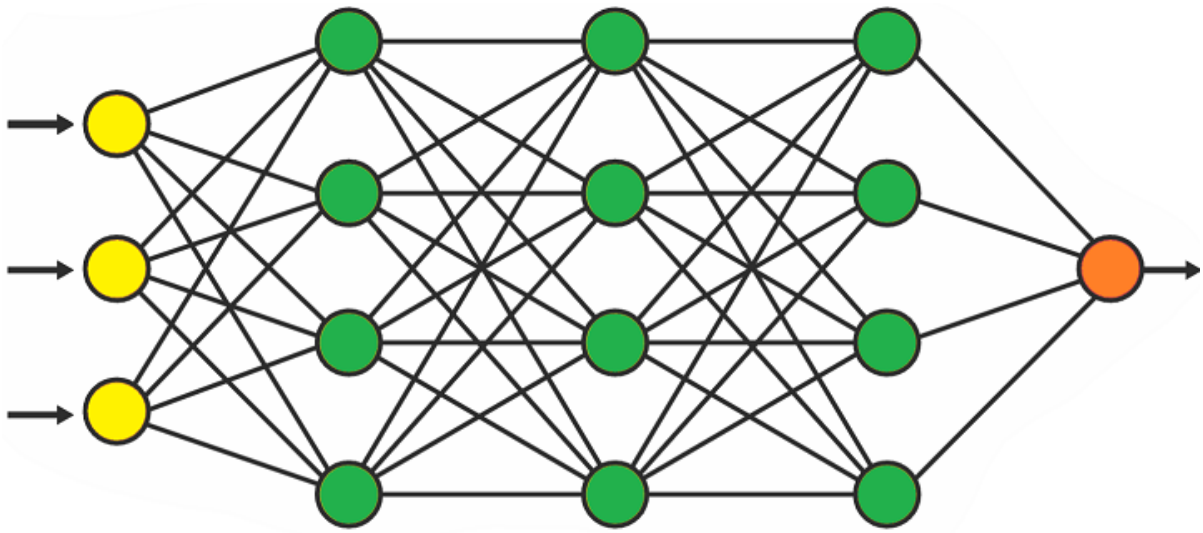
The recursion continues until a stopping criterion is met. Common stopping criteria include reaching a maximum depth, having a minimum number of samples in a node, or achieving perfect purity (Kingsford & Salzberg, 2008). When a stopping criterion is met, the node becomes a leaf node. The majority class (for classification) or the average value (for regression) of the samples in that node becomes the predicted output (Yan-yan & Ying, 2015). The ability to construct simple decision trees and their adaptability make them a popular choice for weak learners in ensemble methods (3.4.7**Error! Reference source not found.**), which are used in the studies reported in Chapters 7and 9.

3.4.3 Multi-Layer Perceptron

The MLP is a type of artificial neural network (Murtagh, 1991). While neural networks are often synonymous with deep learning, the MLP straddles the boundaries of both (LeCun et al., 2015). Whilst not capable of self-feature extraction, it is classed as a deep learning algorithm once the network achieves a certain size or "depth" (Murtagh, 1991). Drawing inspiration from neurons and synapses of the brain, the MLP can develop complex non-linear relationships within data (Brady, 1991), shown in Figure 3-5.

An MLP will comprise three or more layers housing neurons, a minimal multi-layer perceptron includes an input layer conveying features, one hidden layer, and an output layer, as shown in shown in Figure 3-5 (Murtagh, 1991). Notably, MLPs are fully connected, meaning all data from a single neuron in one layer is distributed to all neurons in the subsequent layer (LeCun et al., 2015). The training process employs a technique known as backpropagation (Hecht-Nielsen, 1989). In essence, data are first passed forward through the algorithm, with the ensuing errors in prediction passed back through the algorithm and used to update the model parameters (Janiesch et al., 2021).

Figure 3-5: A representation of a multi-layer perceptron. Yellow marks the input layer, green shows 3 hidden layers consisting of 3 neurons each, with orange marks the output layer. The black paths show the connections between the neurons, through which the data will pass.



Hecht-Nielsen, 1989 defines this process as the following. Considering training data $D = \{(x_n y_n)\}_{n=1}^{N}$, where x_n denotes the features of data instance n, and y_n is the corresponding class label. The maximum likelihood estimation is employed to derive parameter estimates that maximise the probability of observed data. The parameters, denoted as θ, with:

$$\hat{\theta}_{MLE} = \arg max_{\theta} P(\boldsymbol{Y}|\boldsymbol{X};\theta) \quad (3\text{-}24)$$

Where:

$$\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^{L} \quad (3\text{-}25)$$

Rather than maximising a term, the emphasis is on minimising the error term, often referred to as negative log-likelihood (Aitkin & Foxall, 2003):

$$Error_n = -\sum [\![y_{nk} \log P(\widehat{y_n} = k|x_n;]\!] \theta) \tag{3-26}$$

By leveraging the error term, convergence towards optimal parameters is sought. This process is, however, susceptible to local minima, particularly with smaller datasets. Gradient descent is employed to iteratively reduce the cost function by adjusting weights and biases for each data point:

$$w^{(t+1)} = w^{(t)} - \eta \frac{\delta E}{\delta w} \tag{3-27}$$

$$b^{(t+1)} = b^{(t)} - \eta \frac{\delta E}{\delta b} \tag{3-28}$$

This optimisation requires both forward and backward propagation. Forward propagation initiates with randomly set weights and bias terms. As data passes the network, neuron outputs are calculated. Upon reaching the final neuron, data passes through the last activation function, assigning a probability to the data's class. The process proceeds forward, enabling the initial prediction of model performance. In the output layer, a prediction probability $\hat{y}$ for x_i is calculated, facilitating estimates of $\hat{y}$ andE_n. The cumulative error is computed as:

$$E = \sum_n E_n \tag{3-29}$$

Backpropagation computes the gradient of the error concerning weights and biases, guiding updates to minimise error.

MLPs offer a form of deep learning algorithm that is more suitable for smaller datasets, as used within this thesis (Chapter 9), whilst still capable of capturing complex non-linear patterns (LeCun et al., 2015). This is due in part to the lack of deep feature creation, which requires multiple layers and connections to extract patterns, allowing for the creation of a simpler, less deep, network that will be more resistant to over-training (Brady, 1991).

3.4.4 Support Vector Machine

A Support Vector Machine (SVM) is a supervised ML algorithm widely used for classification and regression tasks (Moguerza & Muñoz, 2006). It works by finding a hyperplane in a high-dimensional space that best separates the data points into distinct classes (Hearst et al., 1998). The goal of an SVM is to maximise the margin—the distance between the hyperplane and the nearest data points from each class, known as support vectors (Pisner & Schnyer, 2020). By maximising this margin, the SVM ensures better generalisation and robustness to new, unseen data (Boswell, 2002). SVM classifiers are a popular choice in machine learning and have been used to create state-of-the-art classifiers for HAE detection in specific settings (Wu et al., 2014, 2018).

Hearst et al., 1998 define SVM support vector calculation in the following. Consider a training dataset with m samples (x_i, y_i) where $i = 1,2, \dots, m$. Here, $\boldsymbol{x_i^T} = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ represents a d-dimensional feature of the sample i, and $y \in \{-1, +1\}$ signifies coded labels for each class[9]. The label $+1$ is assigned to positive classes, while -1 is attributed to the negative class. The training data is effectively separated by the hyperplane $w^T x_i + b = 0$, where w is the weight vector and b is the bias term.

The SVM aims to ascertain optimal values for both the weight and bias terms. Once the partition's location is determined, the support vectors, i.e., the nearest points to the hyperplane, are identified. These support vectors play a pivotal role in establishing the marginal hyperplanes, H_1 and H_2, defined as:

$$H_1: [\![(\boldsymbol{w}]\!]^{\boldsymbol{T}} \boldsymbol{x_i} + b) = 1 \tag{3-30}$$

$$H_2: [\![(\boldsymbol{w}]\!]^{\boldsymbol{T}} \boldsymbol{x_i} + b) = -1 \tag{3-31}$$

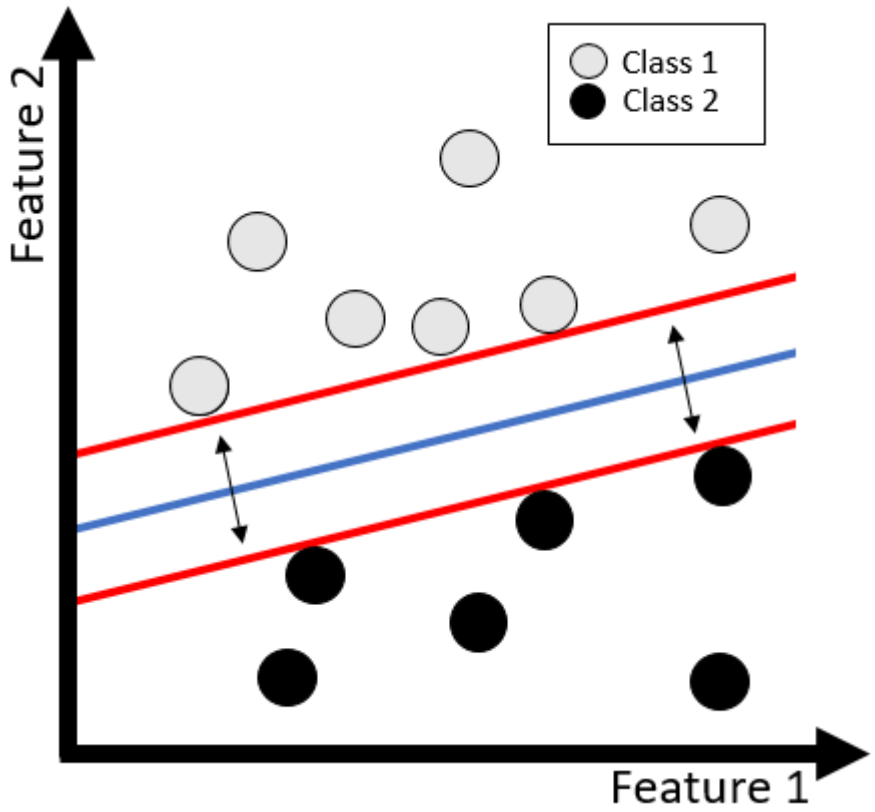
These relationships are visually depicted in **Error! Reference source not found.**. Any point correctly classified satisfies the inequality:

$$y_i [\![(\boldsymbol{w}]\!]^{\boldsymbol{T}} \boldsymbol{x_i} + b) \geq 1 \tag{3-32}$$

For$\boldsymbol{x_i}$, $i = 1,2, \dots, m$, the distance between the separating hyperplane and the marginal hyperplanes is calculated as $\frac{2}{||\boldsymbol{w}||}$, where $\left|\left|\boldsymbol{w}\right|\right| = \sqrt{\boldsymbol{w^T w}}$. This geometric interpretation

underscores the SVM's goal of maximising the margin, providing a robust classification mechanism.

Figure 3-6: A simplified support vector classifier, the classification hyperplane (blue) separates the two classes. The minimum distance between positive and negative hyperplanes are show in red.



The SVM classifier incorporates additional methods to enhance its ability to accurately predict data instances (Mammone et al., 2009). The Kernel Trick is one such technique, allowing SVMs to adeptly handle non-linear decision boundaries by transforming input features into a higher-dimensional space (Noble, 2006). This transformation occurs implicitly, avoiding the explicit calculation of new feature representations and ensuring computational efficiency (Noble, 2006). In scenarios where data is not perfectly separable, the Soft Margin SVM introduces a slack variable, providing a level of flexibility that accommodates misclassification (Noble, 2006). This proves valuable when dealing with noisy or overlapping data, contributing to the classifier's robustness in real-world situations (Boswell, 2002). Moreover, the SVM's effectiveness is further fine-tuned through the C parameter, which determines the balance between achieving a smooth decision boundary and accurately classifying training points (Noble, 2006). A smaller C results in a wider margin but may allow for

some misclassifications, while a larger C seeks accurate classifications at the expense of a narrower margin (W. Zhang et al., 2022). These methods collectively contribute to the SVM's versatility and performance across a spectrum of classification challenges. These easily configured variables lend themselves to scenarios where data is small and not perfectly separable, this helps to create a generalisable model (Roy & Om, 2018; Tabrizi et al., 2020).

3.4.5 Logistic Regression

Logistic regression, a probabilistic binary classification method, is employed to estimate the probability of an event occurring when confronted with new data, where 'logistic' refers to the log of the odds probability (Equation 3-33) (Sperandei, 2014). During the training phase, the model analyses the relationship between features and labels for both classes, predicting a partition between the two variables in a multidimensional feature space (Menard, 2002). Subsequently, when novel data is introduced, the classifier determines its position on the curve (Menard, 2002). This may then be translated it into a meaningful prediction probability of which of the two classes the data may belong to (Hosmer et al., 2013).

$$odds = \frac{P(events)}{1 - P(event)} \tag{3-34}$$

Menard, 2002 defines the underlying methods of the logistic regression classifier as the following. The probability of a data point belonging to class one is defined asClick or tap here to enter text.:

$$\mathrm{P}(y = 1|X) = p(X) \tag{3-35}$$

Where $y = 1$ is the probability of a data point belonging to C_1 after observing the feature vector X. The partition is formed by an activation function, often a sigmoid function which translates input feature vectors into probabilities, bounded between zero and one for predictions:

$$p(x) = \frac{1}{1 + e^{-\beta X}} \tag{3-36}$$

With the general formulas outlined, the purpose of the logistic regression classification algorithm is to predict the parameter vector $\hat{\beta}$ to maximise algorithm performance.

The classifier is optimised so that $\hat{\beta}$ values yield $\widehat{p(x)} \cong 1$ for data from C_1 and $1 - \widehat{p(x)} \cong 0$ for data from C_0. The optimisation of $\hat{\beta}$ is described by the likelihood function:

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - \log\left(1 + e^{\beta x_i}\right) \tag{3-37}$$

The value of $\hat{\beta}$ is determined by maximising this function:

$$\beta = argmax_{\beta} L(\beta) \tag{3-38}$$

As a transcendental equation, $\hat{\beta}$ can be estimated using a solver over multiple iterations. The solver process involves initialisation, creating initial predictions, updating $\hat{\beta}$ values to increase the prediction, and repeated iterations until convergence criteria are met. Once solved, the function $p(x)$ provides predictions for new data. A decision threshold is then implemented, classifying values above as class 1 and below as class 0.

Logistic regression is widely preferred for small datasets due to its simplicity, efficiency, and robustness (Alpaydın, 2013). It performs well under limited data conditions, often outperforming more complex models like neural networks (Arshad et al., 2023). LR's linear decision boundary and few parameters reduce overfitting risk and computational demands (Komarek & Moore, 2003). Its interpretability through model coefficients makes it ideal for applications requiring clear insights into feature importance (Arshad et al., 2023). However, LR's performance can be affected by rare events, multi-collinearity, and nonlinear predictors in small samples (Bergtold et al., 2018).

3.4.6 Ensemble Classifiers

Ensemble methods in ML are techniques that combine multiple individual models to create a more robust, accurate, and generalised predictive model (Sagi & Rokach, 2018). These methods are reliant on the use of a ‘weak learner’ or base model which provides output predictions (Horng et al., 2019). In theory, none of these individual models should provide particularly accurate results, but as an ensemble, the classifiers should perform well (Rokach, 2019; Sagi & Rokach, 2018). These are typically decision tree models but can be used with any model (Rokach, 2019).

Ensemble algorithms are amongst the most popular choices for a variety of ML challenges, with many varieties available (Bentéjac et al., 2021; Rane et al., 2024). They were selected for use in this study due to having previously exhibited state-of-the-art performance in classification tasks, including HAE detection (Goodin et al., 2021). The ensemble methods used within the thesis are:

- Random Forest
- Adaptive Boosting
- Gradient Boosting

3.4.7 Random Forest

A Random Forest is an ensemble learning method that leverages the power of multiple decision trees to improve predictive accuracy and control overfitting (Liaw & Wiener, 2001). Instead of relying on a single decision tree, Random Forest builds a collection or "forest" of diverse trees, and their outputs are combined to make predictions (Breiman, 2001). In a Random Forest, multiple decision trees are constructed using a technique called bootstrap aggregation (T. H. Lee et al., 2020). Bagging involves training each tree on a random subset of the training data, allowing different trees to learn from different perspectives of the dataset (T. H. Lee et al., 2020). The trees within a Random Forest are often termed "weak learners" because their construction prioritises simplicity over precision (Rokach, 2019). Rather than striving for highly accurate models on their own, these trees focus on creating a multitude of diverse classifiers, providing slightly better than random predictions (Horng et al., 2019). The overarching goal is to build a robust ensemble where the collective wisdom of these diverse learners, when aggregated, yields accurate and reliable predictions (Rokach, 2019). This diversity is crucial for reducing the risk of overfitting present in individual decision trees (Sagi & Rokach, 2018). The final prediction of a Random Forest is determined through a majority vote for classification tasks, or an average for regression tasks of the predictions made by individual trees (Horng et al., 2019).

3.4.8 Adaptive Boosting

The adaptive boosting (AdaBoost) method is used to create ensemble classifiers and is particularly effective with weak learner algorithms, such as decision trees (An & Kim, 2010). An & Kim, 2010 define it as the following. The algorithm initiates the data with specific weights, equal for all data points within the training dataset:

$$w_i = \frac{1}{N} \tag{3-39}$$

Where N is the total number of data points. A null classifier, $f_0(x) = 0$ is created for $t = 0$ to T, the number of sequentially created sub-models.

The next step involves creating a training sample from the dataset, considering the training weights. A weak learner, for example, where a decision tree is then trained on this sample. The data is labelled, and classification errors are measured using a loss function. AdaBoost commonly employs the exponential loss function:

$$L(f) = \frac{1}{N}\sum_i e^{-y_i f(x_i)} \tag{3-40}$$

Where N the number of data points in the training is set, y_i is the true label and $f(x_i)$ is the predicted label from the weak learner. The algorithm then calculates a weighted error term e_t for each iteration:

$$e_t = \frac{\sum_{i=1}^{n} e_i w_i}{\sum_{i=1}^{n} w_i} \tag{3-41}$$

The coefficient λ_t is calculated as:

$$\lambda_t = \frac{1}{2}\ln\left(\frac{1 - e_t}{e_t}\right) \tag{3-42}$$

Weighted error term lambda is then required to be calculated, step size for adaptive boosting, for the weights of each data in the model. With this calculated, the weights of the data points are updated. This differs for data points that were calculated correctly and incorrectly. This term acts as a step size for adaptive boosting and influences the weights of each data point in the model. The weights of data points are updated based on correctness:

For incorrect: $w_i \leftarrow w_i e^{\lambda_t}$

For correct: $w_i \leftarrow w_i e^{-\lambda_t}$

This reduces the value of the weight for correctly classified data points and increases it for incorrectly classified data points. Normalisation follows, ensuring the sum of the total weight to one. A new classifier is then created, either by training with a weighted Gini index term or by creating a sample where the classifier is more representative of

the dataset. This is achieved by selecting random values, including the weighting term. Finally, the sum of predictions is used to predict the values for the classifier's outputs.

Since AdaBoost adjusts weights dynamically to focus on difficult examples, it can maximise the learning potential from every available data point, improving classification accuracy even in the presence of sparse samples (Hastie et al., 2009b). Furthermore, AdaBoost does not require tuning complex hyperparameters, which reduces the risk of overfitting that might arise when small datasets are used with more complex algorithms (Sharma & Sharma, 2016). Additionally, the ensemble's additive nature—where new models correct the errors of prior ones—ensures efficient utilisation of the small dataset without the need for extensive data splitting or complex model structures (Raschka, 2018). This focus on incremental improvement is advantageous when large, diverse datasets are not available, though the algorithm remains sensitive to noise or outliers, which may disproportionately influence performance when data is scarce (Sharma & Sharma, 2016).

3.4.9 Gradient Boosting

Gradient boosting classifiers are an ensemble classification method (Friedman, 2002; Natekin & Knoll, 2013). This ensemble is formed by iteratively refining the predictions of a weak learner algorithm based on the class labels of the data and the previous prediction errors. Once this collection of trees has been established, the algorithm provides a final prediction by amalgamating the outputs of all the iteratively improved algorithms. Its ability to sequentially fit weak learners, like shallow decision trees, allows it to capture intricate patterns in the data without requiring extensive pre-processing or feature engineering (Friedman, 2002). Furthermore, gradient boosting incorporates regularisation techniques such as shrinkage and subsampling, which help mitigate the risk of overfitting which is a common concern with small datasets, by controlling model complexity and variance (T. Chen & Guestrin, 2016). The flexibility of the algorithm in optimizing non-linear relationships makes it well-suited for kinematic data, where variables such as velocity, acceleration, and joint angles often exhibit complex interdependencies (T. Chen & Guestrin, 2016). This is illustrated by the use of an XGBoost algorithm in order to detect genuine HAEs in Australian Rules football (Goodin et al., 2021). However, the algorithm’s sensitivity to hyperparameters like learning rate and tree depth requires careful tuning to avoid overfitting, particularly when working with limited data (T. Chen & Guestrin, 2016).

In the initialisation phase of a binary classification gradient boosting algorithm, the first step involves crafting the initial weak learner. This learner generates a uniform prediction for every feature vector, regardless of the specific feature values or class labels. The predicted value, in this context, represents the logarithm of the odds for accurately predicting the positive class, a value consistently applied across all instances during the algorithm's first iteration. This prediction is improved over multiple iterations until a specific ending criterion is met.

The algorithm is trained using a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ where $i = 1$ to n, where x_i represents the feature vector of an instance of i, and y_i is the corresponding class label. A differentiable loss function denoted as $L(y_i, f(x))$, is employed, where $f(x)$ or γ is equal to the logarithm of the odds value. The initial weak learner can be expressed as[20]:

$$f_0(x) = \arg min_\gamma \sum_{i=1}^{N} L\,(y_i, \gamma) \tag{3-43}$$

The primary objective of the classifier is to determine values for the prediction of γ that minimise the loss function. For the first classifier, $f_0(x)$ can be simplified to:

$$f_0(x) = \log\left(\frac{p}{1-p}\right) \tag{3-44}$$

Here, p is calculated as $p = \frac{Y_1}{Y_0}$ with Y_1 representing the number of instances of the positive class within the training data and Y_0 being the instances of the negative class. This formulation sets the foundation for subsequent iterations and improvements in the gradient-boosted algorithm. This process of improvement is repeated for each tree,m in a pre-specified number of trees M, where each instance i of the dataset of size N will be passed through the algorithm.

To improve the function, the original algorithm, $f_0(x)$ is presented with each data point i from the dataset. For each of these N datapoints, the residual, γ_{im}, is calculated and summed to provide the measure of model performance, shown in:

$$\gamma_{im} = -\left[\frac{\partial L\big(y_i, f(x_i)\big)}{\partial f(x_i)}\right]_{f=f-1} for\ i = 1,2 \ldots n \tag{3-45}$$

A second decision tree is constructed to predict the residual values calculated from the original weak learner's predictions, using the dataset features x_i. Where y_i is the observed value, f is the previous prediction. The value of γ used is selected so that it will minimise the residual, making it equal to 0. This process is repeated for leaf, $R_{i,j}$ and each sample relevant to that leaf, is summed, expressed as:

$$f_m(x) = f_{m-1}(x) + v\sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right) \tag{3-46}$$

Based on the previous prediction and the previous residuals, the new prediction is created and updated. Here is introduced the learning rate v, with controls the rate of change of the new predicted values by reducing the influence of the previous predictions.

Now with the updated model, a new prediction is calculated for each sample. The previous prediction is taken using the last model, this is then repeated until the stopping criteria is met. This form the full model, $F_M(x)$, and predictions of the class that a feature vector will belong to can be made, by aggregating the scores of all the weak learners with the following:

$$\hat{y} = \frac{e^{log-odds}}{1 + e^{log-odds}} \tag{3-47}$$

3.5 Algorithms Performance and Analysis

3.5.1 Shapley Additive Explanations

As the use of ML became popular in research, many labelled the internal workings of the ML process as "black box" ideas (Lundberg & Lee, 2017). Whilst powerful models are creatable whilst labelling it as a black box, this does not help to develop an understanding of the value of the features in the process, adding little to domain knowledge. Since then, a group of methods known as "explainable AI" have been developed, aimed at allowing the easy interpretation of ML models. If these methods, one of the more popular is Shapley additive explanations, known as SHAP. This have become a common method of providing objective measures of feature importance, which has a place in feature selection, and adding to domain knowledge with its interpretations.

This method provides an assessment of each feature and how the magnitude of each feature affects a ML models predictions of the dependant variable. More specifically, SHAP values are calculated using an adaption of game theory. SHAP values are calculated by passing the training dataset through the SHAP framework, this assigns a SHAP value to each feature of each data point, and this value represents the contribution of that feature to the model output. The SHAP value $\emptyset_j$ of feature j can be defined as the following:

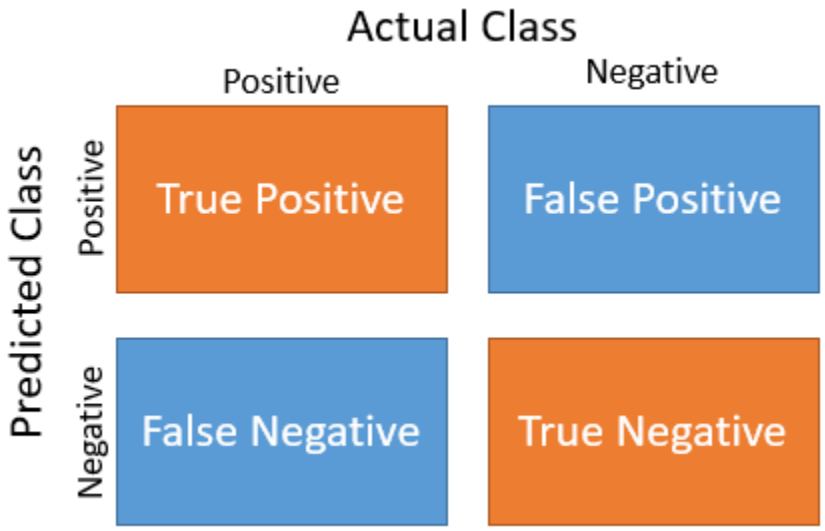
$$\emptyset_j = \frac{1}{|N|!}\sum_{S\subseteq N\{j\}} |S|!\,(|N| - |S| - 1)!\,[f(S \cup \{j\}) - f(S)] \qquad (3\text{-}48)$$

From the following, $|.|$ represents the number of elements in the set, N is the original feature set, and S represents a feature subset from N. Furthermore, $N\backslash\{j\}$ is a subset of all features previous to the feature j, and $f(S)$ is the output of the ML model of the feature subset S. Finally, $f(S \cup \{j\}) - f(S)$ is the cumulative contribution of feature j to the output of the model. As shown in the above equation, the SHAP value of ϕ_j of feature j is estimated by calculating the average contributions in all possible permutations of the subset (Lonini et al., 2018).

3.5.2 Model Performance

In binary ML classification tasks, there are four ways that the label of the data can be predicted: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) (Sofaer et al., 2019). The terms positive and negative refer to which class the model/classifier predicted the data would belong to, whilst the true and false refers to whether the model/classifier was correct in this prediction. These predictions can be represented within a confusion matrix, which reports all of the models predictions, with no consideration of the certainty of each prediction. An examples of a confusion matric is shown within Figure 3-7.

Figure 3-7: An illustration of how machine learning models may correctly or incorrectly predict class labels, known as a confusion matrix.



From the confusion matrix many performance metrics can be derived, with popular metrics including the following.

Accuracy: Measures the overall correctness of predictions made by the model in:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-49}$$

Positive Predictive Value (Precision): This indicates the model's ability to correctly identify the positive instances out of all instances predicted as positive.

$$Precision = \frac{TP}{TP + FP} \tag{3-50}$$

Recall (Sensitivity or True Positive Rate): This measures the model's ability to correctly identify positive instances out of all actual positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{3-51}$$

Specificity (True Negative Rate): This measures the model's ability to correctly identify negative instances out of all actual negative instances.

$$Specificity = \frac{TN}{TN + FP} \tag{3-52}$$

F1-Score: Represents the harmonic mean of precision and recall, providing a balance between the two metrics.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3-53}$$

False Positive Rate (FPR): Indicates the proportion of actual negative instances that are incorrectly predicted as positive.

$$FPR = \frac{FP}{FP + TN} \tag{3-54}$$

False Negative Rate (FNR): Indicates the proportion of actual positive instances that are incorrectly predicted as negative.

$$FNR = \frac{FN}{FN + TP} \tag{3-55}$$

Metrics from the confusion matrix are point-based metrics that focus on specific aspects of the model's performance, evaluating predictions across different classes directly. Classifiers may return a probability of an instance belonging to a class rather than an actual class label. Probabilities will be turned into labels depending on which side of a decision boundary they fall on, often set at 0.5. By changing the decision threshold of the classification model, the metrics of the same classifier may be altered as labels could change depending on the original prediction probability. Therefore these metrics may not present a full representation of the model's ability to classify data.

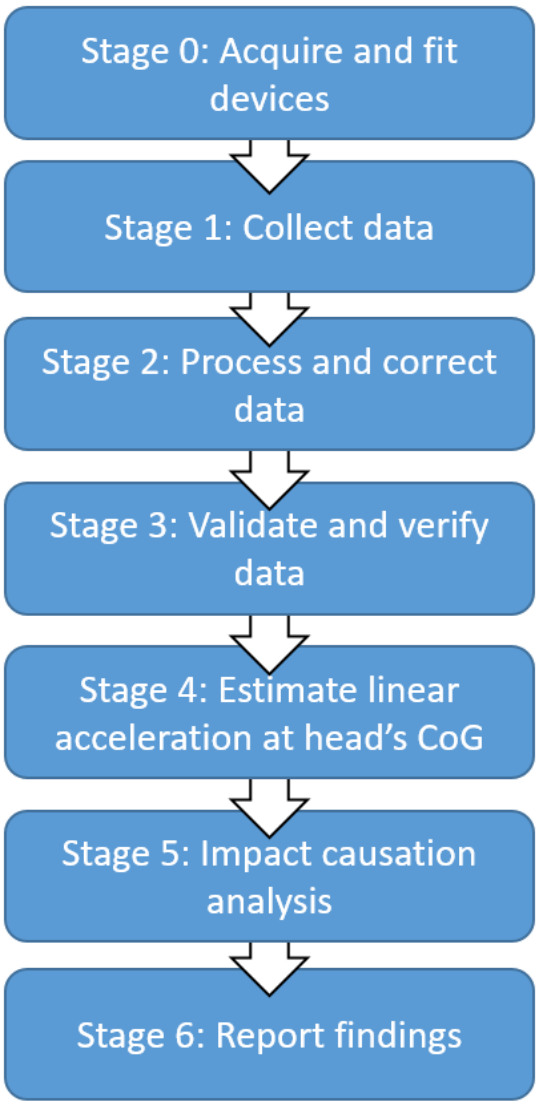
Curve-based metrics provide an overall assessment of a model's performance across different thresholds for binary classification tasks (Sofaer et al., 2019). They are computed by varying classification thresholds, generating curves that show how performance changes as the threshold moves. This provides a more comprehensive view of how completely the classifier has split the data. Two such examples of curve-based metrics are AUROC (Area Under the Receiver Operating Characteristic Curve) and AUPRC (Area Under the Precision-Recall Curve). They consider the trade-off between true positive rate and false positive rate (AUROC) or precision and recall (AUPRC) across a range of classification thresholds. AUROC and AUPRC are less affected by class imbalance and can still provide valuable performance information even when the dataset is imbalanced. AUPRC is particularly useful when dealing with highly imbalanced datasets or when the positive class is more important than the negative class. Metrics from a confusion matrix offer more straightforward interpretability, AUROC and AUPRC are comprehensive but less intuitive; they

provide an overall view of the classifier's performance but might be harder to interpret directly.

4 Moving from Review to Methods and Results

This chapter concludes the review of the literature, which has provided a foundational understanding of the methods that appear in the following chapters. In these following chapters, the methodologies and findings from the studies conducted are detailed. The following chapters are ordered to best follow the pipeline illustrated in Figure 4-1. This pipeline represents the steps that an IMG user may undertake in order to make use of the data gathered by IMGs, this is explained further in the following paragraphs.

Figure 4-1: A pipeline illustrating how data goes through processing from collection to reporting.



Step zero is the actual acquisition of suitable devices and undergoing the relevant device fitting process. Once devices have been fit to the users, step one, data collection

from players participating in training or matches may occur. Step two is data processing, transforming data from the raw signals to a usable form, including filtering and estimation of rotational acceleration from rotational velocity. Step 3 is the verification/validation of data, where the reliability of the data and device is assessed. By including this stage as early as possible it means that data can be excluded from later steps, reducing the resources required to complete the pipeline.

The order of steps 4 and 5 could be interchanged depending on the study outcomes. The ordering in Figure 4-1 was selected as "Stage 4: Estimate linear acceleration at the head's CoG" is a step all studies should undertake in order to report accurate linear acceleration, whereas not all studies will investigate HAE causation. This also presents an opportunity to exclude data from the study if the value of the linear acceleration at the head's CoG falls below a specified threshold. Step 5 is determining the causes of the impact that led to the recording, with the final stage, 6, being reporting the data.

In this thesis, Chapter 5 covers how steps 0-3 the collection and preparation of data which is then used in Chapters 7, 7, 8, and 9. Chapter 0 details processing the data gathered in chapter 5 in order to be used by ML algorithms in Chapters 7 and 9. Chapter 7 uses the data collected and processed in the previous two chapters to create a ML algorithm that can automate "Step 3: Validate and verify data".

Chapter 8 head measurements to assess the importance of head size when reporting linear acceleration gathered by IMGs. The HAE data used in this chapter is the same data gathered in Chapter 5, though a separate cohort is used to collect the head size measurements. Chapter 9 uses the data from chapters 5 and 0 in order to automate step 5 of the pipeline, impact causation analysis. While the final chapter, **Error! Reference source not found.**, presents the specific challenges, future directions, and conclusions from this thesis.

5 Collection, Processing, and Validation of Head Impact Data Introduction

While researchers have extensively used IMGs in HIT research, studies have shown devices can record spurious events, such as shouting or device removal/insertion (E. M. P. Williams et al., 2021; Wu et al., 2014). To account for this, most HIT devices use algorithms to remove spurious recordings from the dataset (Patton, Huber, Jain, et al., 2020). Despite the use of algorithms to remove spurious events, studies with on-field data collection have shown that HIT devices may still record false positives events (Patton, Huber, Jain, et al., 2020). Because of this, video verification is considered an important step when collecting HIT data with IMGs, to ensure the quality of the data (Patton, Huber, Jain, et al., 2020). Video verification is reviewing the IMG recordings along with video footage to confirm that recorded data correspond to a HAE, allowing for the removal of spurious data from the dataset.

This chapter reports the collection, validation, and verification of IMG data and the collection of video footage, corresponding to steps 0 – 4 in Figure 4-1. It is also discussed how the methods used relate to the best practices described within the CHAMP documentation, and the challenges and limitations associated with data collection (Arbogast et al., 2022). This will allow us to assess the accuracy, reliability, potential sources of bias in the data collected and the implications of the use of these data for athlete safety and research.

5.1 Methods

5.1.1 Participants

Data were collected from female collegiate RU players competing within the British universities and colleges premier division south, during the 2021/22 RU season. The work was conducted within the framework outlined in the CHAMP 2022 project to ensure best practices. A total of 25 devices were issued to players within a UK university women's RU squad. Participants were included in these studies if they trained regularly and were expected to feature in the first team in competitive matches during the 2021/22 season. The devices issued to the players remained functional during the six competitive matches recorded, with the data collected from players representing all positions. Before the beginning of the study, institutional ethics approval was obtained from the Swansea University College of Engineering Research

Ethics and Governance Committee. In compliance with the ethical approval (ethical approval number FP_01-10-21), all subjects provided informed consent to their inclusion in the study. Anthropometric measurements were taken from players before the season, while measures of neck strength were taken before and throughout the season. A survey was completed by each player in pre-season to provide details of their sporting background and injury history.

5.1.2 HIT Devices

Data were collected using the Prevent Biometrics Hybrid IMGs (Prevent Biometrics, Edina, MN, USA), which have been previously validated both in laboratory testing and on-field (Bartsch et al., 2014; Kieffer et al., 2020). All IMGs contained a 3.2 kHz three-axis angular rate sensor, three 3.2 kHz single-axis linear accelerometers, a 130mAh battery, microcontroller, proximity sensor, internal storage of up to 460 events and BLE transmission (Bartsch et al., 2014). The angular and linear sensors had measurement ranges between ± 35 $rads^{-1}$ and ± 200 g, respectively. Full descriptions of the technical specifications of the IMGs have been reported by (Bartsch et al., 2014; Hedin et al., 2016; Kieffer et al., 2020; Miller et al., 2018).

A recording was triggered when a single axis exceeded a 3 *g* linear acceleration. This threshold could be altered within the Prevent Biometrics platform to include or exclude recordings of varying magnitudes from the dataset. Once triggered, a segment of kinematic data beginning 10 ms before the limit being exceeded, and +40 ms after was recorded. These data were temporarily stored temporarily on the IMG before being transmitted via an iOS tablet on the touchline via Bluetooth. Once the data were transmitted to the iOS tablet, data were uploaded to the Prevent Biometrics cloud server.

The data transmitted to the Prevent Biometrics cloud server was fed to an algorithm to determine the cut-off frequency for the Butterworth filter that would be used to filter the recording. Prevent Biometrics' cut-off frequency estimating algorithm would then select an appropriate digital filter depending on the noise in the recording which could be either 200, 100 or 50 Hz (Tooby et al., 2022b). Higher levels of noise would result in the ML algorithm selecting a lower frequency filter cut-off (2.4.5). The purpose is to reduce the likelihood of overestimating the impact magnitude by inflating the value with high-frequency signals.

In addition to the recorded rotational velocity and linear acceleration, rotational acceleration was numerically estimated and reported using the rotational velocity. Data were then subject to translation such that values would be reported from a 50th percentile male head CG. Prevent Biometrics also use an additional proprietary algorithm to improve the accuracy of the recording which was developed following the laboratory validation of the device. Data were reported so that the axis would align with the standard human axes as reported within the Instrumentation for Impact Test - Part 1 - Electronic Instrumentation J211/1_201403 documentation, this is illustrated in Figure 5-1. The standard covers the different types of sensors, signal conditioning equipment, and data acquisition systems that are used to measure and record various parameters during impact testing, such as acceleration, force, displacement, and deformation.

An infrared sensor and emitter are contained within the Prevent Biometrics IMG devices, which are used to measure the closeness of fit between the device and the teeth, referred to as the device's proximity sensor. The sensor would return time series data of arbitrary units that reflected the strength of the signal recorded by the proximity sensor. This ranged from 0, indicating no recording, to 999 which was the highest value possible. Impacts that recorded a peak proximity sensor below a given threshold were excluded from the true positive dataset, as were events that contained large changes in the recorded proximity value. The threshold value could be set before the data recording session. These recordings were classified as false positive events and were automatically removed from the true positive recordings. The reasoning for this is that recordings containing low or inconsistent proximity values may have occurred off teeth or the device may have been poorly coupled to the teeth during impact recording.

It is suggested, however difficult to prove, that issues with recording impacts are related to the fit of the device, therefore ensuring that the device fits well from the first session is important to keep the recordings at a high quality (Luke et al., 2024). Continual assessment of the device fit is also valuable as they may deform over time. This will potentially cause issues with the reliability of the device. Retaining positive relations with the participant helps to maintain the use of the devices and sort issues, such as fit or non-device use as they become represented. Ensuring that participants

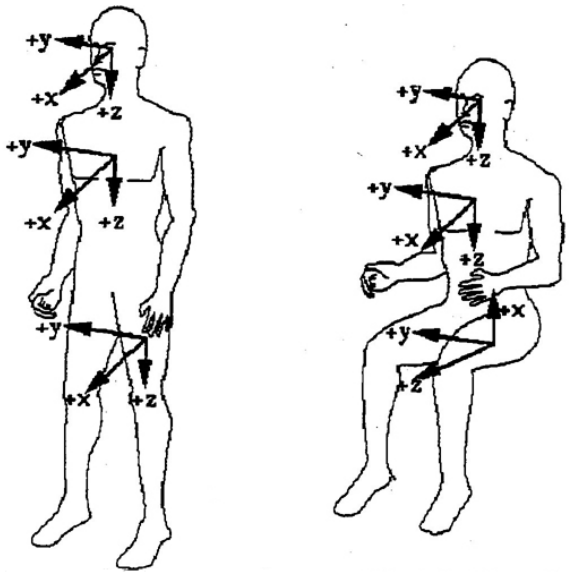
are happy with the devices and provided them on game day will encourage use and allow for more data to be gathered throughout the season.

5.2 Data Preparation

To remove potential sources of bias in following studies, raw data were provided by the company in addition to the filtered datasets. These included all events recorded by the IMGs during the recording sessions with a linear acceleration of > 3 *g*. The first processing step was the digital filtering of the raw kinematic data using a 200 Hz, 4th-order Butterworth low-pass filter. Thus, the cut-off frequency was selected as it is of the same magnitude as filters commonly implemented in conjunction with IMG data processing algorithms. Rotational acceleration was then estimated numerically from the rotational velocity, using the four neighbours of a central point to estimate the acceleration. This method is known as the central point differential method.

The magnitude of each group of kinematic measurements was estimated by finding the Euclidean norm. Event magnitude was not only used for developing characteristics of events but also as an inclusion criterion for recordings. Events were only included within the datasets should they have a peak linear acceleration magnitude of > 9 *g*. This value was selected to be in line with the standard values within the field and greater than the 6 *g* limit of typical voluntary human motion (Rowlands & Stiles, 2012). In line with the Prevent Biometrics proximity filtering system, where events with low or inconsistent proximity sensor values were removed from the dataset. This limit was selected by investigating the typical limits used by Prevent Biometrics.

Figure 5-1: Diagram showing the reporting axis for devices according to "Instrumentation for Impact Test - Part 1 - Electronic Instrumentation J211/1" (Society of Automotive Engineers, 2003).



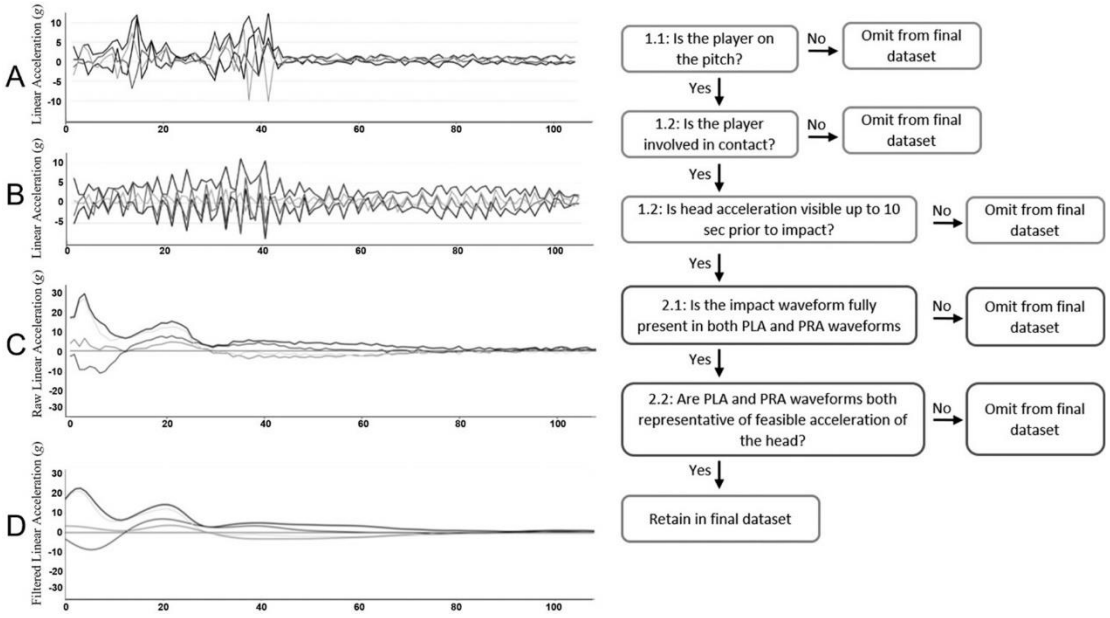
5.3 Impact Verification

For each rugby match recorded, time-stamped videos were recorded from the centre line of the pitch with a device of minimum specifications of 1080p, and 30 fps to enable the identification of events. Should a secondary video camera and operator be available, video footage would be taken from the opposing touchline to provide an alternate angle for impact verification. Each recording was independently verified by two of the co-authors with 18 months of impact reviewing each, using a two-stage verification method. The first exclusion criteria were whether the players were involved in the match at the time of recording. Should players not be included in play at the time of recording then any events during this period could be classified as off-field events and be excluded from the final dataset. To calibrate timing across the video and IMG data, two high-certainty HAEs were found via video review and used to measure the time difference between recording and video time. The time lag, if present, was assumed to be the same across all IMGs. A list of all events with corrected times was then compared to the video footage to verify the HAEs. Events would pass to the next stage if the player wearing the device that recorded the event was seen to be involved in a contact event with a clear head acceleration within a ± 3-second window of the recorded time.

The HAE recording that progressed past this stage then underwent visual inspection. This consisted of assessing the acceleration to determine whether it was representative of the on-field head acceleration. The data used to create these guidelines were developed using data collected in a pilot study in support of the study Williams et al., 2021. Williams et al., 2021 reports that in this pilot work, IMGs were given to a sub-sample of five available participants who completed several activities to generate both HAEs and spurious recordings. This was conducted to create qualitative methods of classifying impacts. These recordings all contained linear acceleration of greater than 10 *g*, using a device of similar specifications to the Prevent Biometrics IMG. This resulted in the collection of 160 biting events, 120 mouthguard removals, 120 mouthguard insertions, 50 of sucking on the devices, and 75 rugby tackles with hit shields. Additionally, many recordings of the participants shouting whilst wearing the devices were also collected. These recordings allowed for the characteristics of the true and false positive events to be greater understood as the characteristics of false positive impacts were studied and identified by comparing the events to genuine, ground truth events. This knowledge was then applied to the current study and used to include or exclude impact.

Figure 5-2 shows characteristic waveforms from a bite (Figure 5-3 (A)), a shout (Figure 5-3 (B)), and a true non-filtered (Figure 5-3 (C)), or filtered (Figure 5-3 (D)) head-impact event. Impacts were excluded from the dataset if waveforms were not representative of feasible head acceleration or if incomplete waveforms were recorded, where data had been dropped during transmission. Should events closely resemble A or B, for example, peak duration of >5ms and multiple peaks forming within 10ms, data would be excluded.

Figure 5-3: The criteria used to assess waveforms recorded from IMGs as reported in Williams et al., 2021. Waves A and B represent spurious signals, whilst C is a genuine signal pre-filtering and D is post-filtering. Originally from "Sex differences in neck strength and head impact kinematics in university rugby union players" by Williams, E.M.P., Petrie, F.J., Pennington, T.N., Powell, D.R.L., Arora, H., Mackintosh, K.A., and Greybe, D.G., originally published in the European Journal of Sport Science, 22: 1649–1658 (2022), https://doi.org/10.1080/17461391.2021.1973573. Used under the terms of the Creative Commons Attribution license (CC BY). For more details, see https://creativecommons.org/licenses/by/4.0/.



Events passing the verification process were labelled as positive events, whilst any on-camera events that failed to pass were labelled as negative events. After impacts were verified as genuine, they were added to a library of genuine impacts. Impacts that could be verified as spurious, or failed at any stage of the verification process were excluded from the genuine impacts and included in the spurious event category. This led to the creation of a genuine and spurious library for each of the six matches participated in through the 21/22 season.

In addition to the genuine or spurious label, more information was recorded about each for the genuine activities. This included the action the player was conducting before impact, which was coarsely described as a ‘ball carrier’ for all events where the player had the ball within their possession or a ‘tackler’, where the player did not carry the ball and was involved in a tackle event. A second category was included which described the mechanism of contact that led to the recording of the head acceleration. This was defined broadly as a primary impact, where there was clear and direct contact to the head, or secondary, where an impact to the body led to the head acceleration

without any DHC. For the impact type study labels of '0' were assigned to NDC and labels of '1' were given to DHCs. For the action recognition study, a label of '0' was assigned to tacklers and labels of '1' were given to BCs. Events that that did not fit the binary categories of tackler or BC were not labelled with the labels of one or zero, however, they remained within the true positive dataset. Should multiple events be recorded in a contact event window, only the first recording was retained. Additional information about each event was recorded adding further details to the impact type, for example, the object the head collided with in primary impact and the direction of impact.

5.4 Results

In the six matches recorded throughout the 21/22 BUCS rugby season, the match squads contained between 12-22 players wearing an IMG. During this time, over 7,500 player match minutes were monitored for HAEs.

5.4.1 Data for Impact Verification

In total, over 10,000 raw data recordings from within these games were provided by Prevent Biometrics. Most of these recordings were excluded from the study due to low linear acceleration magnitude and/or low or inconsistent proximity sensor readings. A total of three impacts were removed from the genuine impact group having failed visual inspection. After these exclusionary steps, a total of 214 HAEs which were confirmed as genuine after video review. A total of 466 recordings were confirmed as spurious during the video verification process, including events such as bites, shouts, insertions, and removals. Data were divided such that five of the six matches were included in the training data, containing 166 head accelerations events and 400 spurious events. The remaining match was used as test data, this contained 48 genuine events and 66 false positive events.

5.4.2 Data for Automated Epidemiology

Of the 214 events determined to be genuine, labels were assigned to 123 impacts for the impact type study, including 91 direct impacts and 46 indirect impacts. The remaining 91 impacts were either unidentifiable or occurred as part of multiple recordings from one contact event. Data were split to create training and test groups of 91 impacts (68 direct and 36 indirect), with a test set of 33 recordings (23 direct, 10

indirect) respectively. This split meant that four of the six matches were used as training data whilst the remainder were used as test data.

For the action recognition study, 169 events were identified for use in further analysis. These included 65 BC events and 104 tackler events. The data were split so that four games containing 135 events appeared in the train set (53 BC, 82 tackler events), whilst 34 events (12 BC, 22 tackler) were used in the test dataset. In all, 45 events were not included in the genuine data as they did not fit the binary classification system.

5.5 Discussion

This work aimed to collect a suitable quality and quantity of data to enable ML studies to be conducted. These data were used to create three datasets, one of which was used to train the impact detection classifier and two for the action recognition studies. These datasets represent some of the few collected for study from female cohorts or RU (Patton, Huber, Jain, et al., 2020).

Data were collected from 25 female participants competing in six matches in an inter-university competitive RU league. Not all players competed in all matches, with the number of instrumented participants in match day squads ranging from 12 – 22. This resulted in a dataset containing 680 impacts that were verified as genuine or spurious events for the genuine impact study. Out of the 680 verified events, 214 were deemed genuine, which meant an incidence rate of 26.8 per match hour from the six matches. Comparatively, incidence rates of HAEs >10 g have been recorded in elite men's rugby of 22.7 and 13.2 per player hour for forwards and backs, respectively, while in elite women's rugby, it was 11.8 and 7.2 per player hour (Tooby, Woodward, et al., 2024).

Given the number of players within this study, perhaps more accelerations could have been expected. This discrepancy in HAE incidence rates could arise from differences in training for the athletes, data processing steps, or lack of camera angles within this study. The collegiate athletes in this study are likely to be less powerful than the elite athletes which may lead to less frequent head accelerations due to lower contact speeds. The differing data processing may also explain the large differences in incidence rates. For example, a linear acceleration threshold of 10 g at an estimated 50^{th} percentile male head CoG was used in the comparative study, as opposed to 9 g at the sensor in this study (Tooby, Woodward, et al., 2024). Further to this, some

potentially genuine events were excluded in this study due to a lack of clear camera angles. The comparative study also suggests that due to the recording biases of IMGs their reported rates are still likely to be underestimates of the true occurrence rates of HAEs in RU. This could have been mitigated by using a rotational acceleration threshold, rather than linear acceleration. This is not typically used by IMGs as rotational acceleration must first be calculated, which requires computational power and makes it more difficult to capture multiple events, but this metric might be more useful once events have been recorded .

For the study of genuine HAE classification dataset study, a dataset of similar scale to those used in previous literature was collected from six matches (Gabler et al., 2020; Kuo et al., 2017). The results from these studies suggest that datasets of this size are effective for developing HIT classifiers. While some studies have utilised significantly larger datasets (Domel et al., 2021; Goodin et al., 2021; Raymond et al., 2022), cross-validated testing has demonstrated that the data from the smaller studies are both robust and reliable, supporting their use for HIT research (Gabler et al., 2020; Kuo et al., 2017).

For the impact causation studies, two datasets were created: one with 123 HAEs verified for impact recognition and another with 169 HAEs verified for action recognition. The volume of data collected is comparable to that of similar studies using alternative data sources, as shown in Table 5.1. With these reliable, validated classifiers, there is potential for the development of similarly robust and high-performing classifiers based on this data. Limitations may be found in the fact that whilst the dataset size is comparable, the classification challenge is different and perhaps more complex in the causation analysis, which would suggest that the performance of the classifiers may be worse. However, this work could serve as feasibility studies for further advancements in the HIT field if highly skilled classifiers are not developed from this data. This could encourage future research to enhance IMG functionality, ease the challenges of causation analysis, and generating practical, data-driven recommendations.

Table 5-1: Dataset sizes within similar, previous studies.

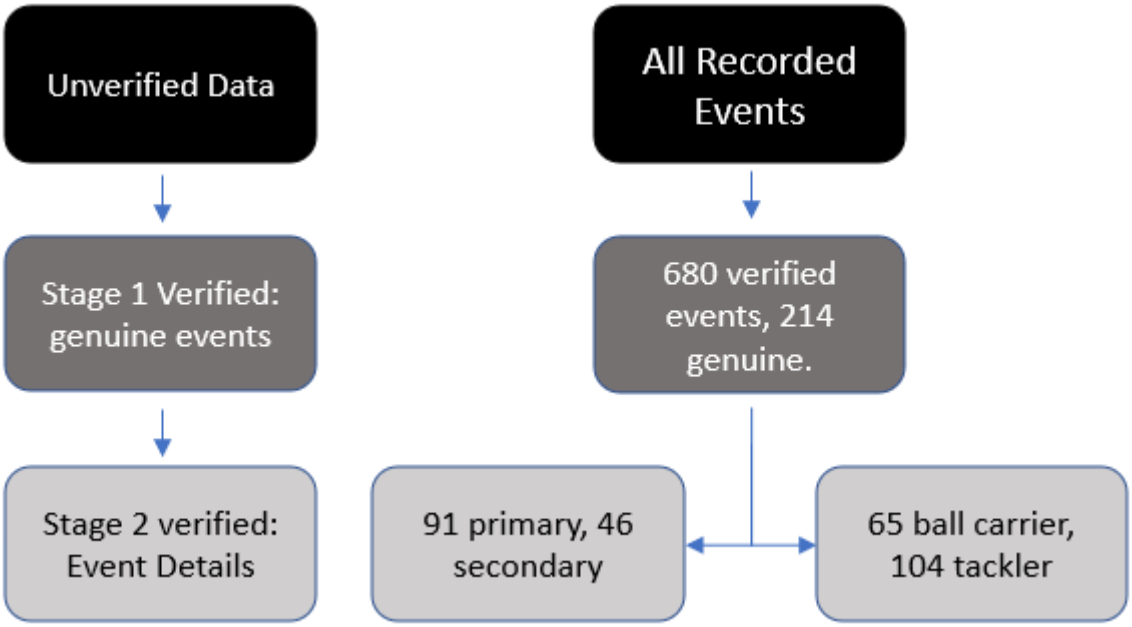
Study	True Positive	False Positive	Total Recordings
(Wu et al., 2018)	156	231	387
(Gabler et al., 2020)	185	379	564
Genuine event detection study	214	466	680
(Raymond et al., 2022)	1,024	10,990	12,014
(Goodin et al., 2021)	1,651	12,059	13,710

The total number of impacts reduced at each stage of the verification pipeline is shown in Figure 5-4. This resulted in different numbers of verified genuine events used within the impact causation studies versus the impact verification study. There was loss of 77 events for the DHC/NDC study and 45 events for the tackler/BC study. Event loss between the layers of verification may occur for several reasons. Events that are recorded from actions which do not fit the binary classification of the tackler vs BC study were excluded from the study, such as event from rucks or mauls. Events where DHC or NDHC could not be established due to a lack of clear footage were also excluded, an issue that may have been improved with more video footage.

A solution to improving the quality of footage available to verify events is by increasing the number and quality of the cameras available to record the match footage. More camera angles, or higher quality footage could give the benefit of uncovering more information pertaining to the HAE. Specific guidance or best practice has not been established in this area regarding these factors. Early guidance has been described within CHAMP documentation, although other than more cameras, greater film quality, and camera angles >2m, no specific recommendations are made (CHAMP ref). Each suggestion comes at some cost. For example, to get more camera angles, more cameras with operators are required. Ball tracking cameras remove the

need for many operators, but come at large financial cost and may not have the reach to capture suitable video from each area of the pitch. Static cameras could be used, but may not capture all of the relevant parts of play. Therefore during study design the feasibility of each option should be assed in order to maximise the ability to record impacts.

Figure 5-4 The stages of impact verification.



The work detailed in this chapter has resulted in the creation of three curated datasets that form the foundation for subsequent analyses in this thesis. These datasets will first undergo a feature transformation, as described in Chapter 6, enabling their application in ML studies. One dataset, containing confirmed genuine and spurious recordings, will be utilised in Chapter 7 to train a classifier capable of automatically distinguishing between these types of recordings. Additionally, two datasets derived from the genuine recordings, flagged with impact causation details, will be employed in Chapter 8 to evaluate the feasibility of training classifiers to detect the causes of impacts. Finally, genuine recording kinematic data will be used in Chapter 9 to investigate the relationship between head size and the estimated linear acceleration at the head's CG. Together, these datasets represent a significant step toward advancing the understanding of head impacts through ML and quantitative analysis.

6 Methods and Results: Feature Development and Extraction

6.1 Introduction

The objective of this thesis is to improve the reliability and functionality of IMGs, specifically in the context of female RU. ML algorithms can support these objectives by processing and interpreting data in ways and speeds that humans are not capable of (X. W. Chen & Lin, 2014). Depending on the type of ML algorithm, different data preparation methods will be required to transform the data into a suitable form for the algorithm (Zheng & Casari, 2018). Chapter 7 and 9 use classical ML algorithms, requiring the data to be transformed into descriptive features (Zheng & Casari, 2018). Section 1.6 discusses the rationale for using this type of algorithm. These features may need to be derived from scratch where domain knowledge is poor, but similar research will provide the opportunity to adopt and adapt features in support of specific goals (Dang et al., 2020).

This thesis addresses three specific classification challenges:

1. The detection of genuine head impacts (Chapter 7)
2. Direct head contacts vs no direct contact head acceleration events (Chapter 9)
3. Ball carrier events vs tackler. (Chapter 9)

Several papers have now been published reporting the creation of head impact detection algorithms, reporting results comparable to human performance in specific settings (Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2014). These studies have used similar feature sets to create these classifiers, predominantly consisting of wavelet, PSD, pulse parameters, and general statistic measures (Wu et al., 2018)(Chapter 3). The frequencies used for these features have ranged from 10 Hz to the cut-off frequency of the digital filter used in data preparation, typically 200 - 300 Hz (Wu et al., 2018). Because these studies have been successful, researchers have a relatively well-defined set of features to use for training the classifier. It remains unclear how the importance of these features varies across different classification tasks.

Action recognition or impact type detection is a task that has yet to be attempted using HIT devices. However, there are numerous publications documenting the use of ML algorithms to predict actions from other wearable kinematic sensors in similar sporting settings (Hendry et al., 2020; Kautz et al., 2017; Tabrizi et al., 2020). A comparative study reports an algorithm trained to detect ruck and tackle events in elite RU players, using video-verified kinematic data and a wearable worn on the upper back (Chambers et al., 2019a). The trained random forest classifier achieved 100% accuracy when tested upon the validation set. This provides a sound body of research with which to build feature sets to best address this classification challenge.

Similarly, no classification tasks have yet focused on detecting DHC (direct head contact) versus NDC (non-direct contact) head acceleration events (HAEs). However, due to the similarities between this classification task and other challenges in the domain, the features established in those related contexts are likely to be valuable here as well. Specifically, the features used in the genuine versus spurious classification challenge, which aim to capture the characteristics of head acceleration, may also provide key discriminating factors for this task. While these features form a firm foundation, the patterns learned may need to be more detailed and sensitive to distinguish effectively between the differing types of head acceleration, reflecting the nuanced differences between DHC and NDC events.

This chapter aims to develop a comprehensive feature group capable of training classifiers to achieve state-of-the-art performance in the genuine versus spurious classification task, while also providing a foundation for benchmarking results in the other classification tasks. By synthesising and adapting feature sets from published work, this research seeks to identify features that can effectively train the classifiers. The methods used to create these features are documented in this chapter, alongside the techniques employed for feature selection and dimensionality reduction, ensuring that only the most informative and valuable features are retained. The resulting datasets, once transformed into features, will then be used to train the classifiers in Chapter 7 and 9.

6.2 Methods

After the data were recorded, they were filtered and transformed to enable their conversion into features vectors with which to train ML algorithms. Five key feature groups are described below and how they were generated.

Pulse Parameters

The prominence, width, and number of pulses were identified from local maxima in the signal. The prominence of each peak was measured by calculating the vertical distance between the highest point and the lowest contour line. The width of each peak was measured by calculating the horizontal distance at the lowest contour line. In the event of there being multiple instances of either measure, the maximum value calculated would be used. The final measure was the total number of peaks per signal. These pulse parameters were measured using the Scientific Python (SciPy) library "find_peaks" (Virtanen et al., 2020).

Positional Derivatives

The first and second derivatives were calculated from each of the kinematic measurements. The first derivative was calculated from the change in two sequential recorded values, with the second derivative calculated by the same process from the first derivative. The maximum absolute value of each signal was used, this providing the maximum first and second derivatives of each waveform. The maximum value of the first and second derivative was found using the NumPy method "diff" (Harris, Millman, van der Walt, et al., 2020).

Power Spectral Density

PSD describes the power of a signal in frequency components (Solomon Jr, 1991). This was calculated in 20 Hz windows using Welch's method between 20 and 200 Hz, with the upper bound determined due to the filter's cut-off frequency (Solomon Jr, 1991). The power values for each frequency were used as a feature, providing ten features per vector. Power spectral densities were found using the SciPy signal package, with the built in Welch's method using a flat-top window (Solomon Jr, 1991; Virtanen et al., 2020).

Wavelet Transformation

A CWT was used to provide time dependant frequency analysis (Rioul & Duhamel, 1992). A CWT was conducted for each signal using the Ricker wavelet function

between 10 and 200 Hz, in 10 Hz increments (Ryan, 1994). Values were taken from a maximum of three time points in the waveform for each kinematic measure transformed. These were at the time points of the maximum absolute value of that particular waveform and the start of the recording. These features were calculated using the Python library PyWavelets (G. Lee et al., 2019). The measurement taken at the start of the recording was not include for the genuine impact classification algorithm, as it was designed to attempt to detect more information about the events leading to the impact.

Statistical Features

Four other features types were used, the first three were the mean, standard deviation, and the area under the curve which was estimated using the trapezoid method. These were calculated from singular kinematic waveforms using NumPy methods (Harris, Millman, van der Walt, et al., 2020). The feature remaining was the principal component, calculated using principal component analysis (PCA) (Wold et al., 1987). This is a dimensionality reduction method that identifies the directions in which the data varies the most (Wold et al., 1987). This was calculated using the Scikit-Learn package (Pedregosa et al., 2011). These features were included only for the action recognition and head impact type classification tasks.

6.2.1 Feature generation, storage, and naming convention

In python, a dictionary was created containing the details for the features that were going to be created. In this dictionary, it contained the name of the feature, the kinematic measure to be transformed and the frequency range of interest should it be relevant. This dictionary was created to provide a list of transformations to an object, which also contained the data. When initiated the object performed a series of transformations on each impact stored within a designated folder, returning a dataframe that contained feature vectors representing each impact. This dataframe was then saved as an excel file so that it could be directly loaded into Python sessions. The code used for feature extraction is shown in the appendix along with part of the output feature CSV file.

Features were given names within the dataframe so they could be referenced in analysis and their impact could be understood. Each name consisted of up to four components. The leading part of the name described the method used to transform the

data, this included PCA, Mean, Std Dev, AUC, WT, PSD, $1^{st}/2^{nd}$ derivative, and Pulse Width/Number/Prominence. The next aspect describes the direction and sensor, with the axes of the sensor represented with Cartesian coordinates and the kinematic measure being L for linear acceleration, A or RA for rational acceleration, and V or RV rotational velocity. Finally, should a number be contained within the name it will correlate to the frequency in Hz that the feature was derived using.

Once calculated, these were appended into a list then added to a dataframe. Once completed a column was appended to the dataframe containing the label values for use by the classification. This was then exported as a dataframe.

6.2.2 Feature Normalisation

Feature normalisation is an important pre-processing step in ML, in which the features are transformed into a similar range of values (Patro & Sahu, 2015). The purpose of which is to ensure that no feature dominates the others in the training process, as some classifiers such as SVM may encounter skewed results when some features with larger ranges are used in comparison to those with smaller ranges (Patro & Sahu, 2015). This can be achieved with several methods, the most common data normalisation methods include Min-Max scaling, Standardisation, Z-score normalisation, and Log transformation (Patro & Sahu, 2015).

Min-Max scaling, which was selected for this project, rescales the values of the features between 0 and 1 (Patro & Sahu, 2015). To scale the data in this way, the minimum value from each feature group is subtracted from each feature, with the resulting values then divided by the range of the feature group (Patro & Sahu, 2015). Min-Max scaling is a simple and commonly used method, capable of dealing with data with skewed distributions. However, its primary drawback is that it is not capable of dealing with outliers well and may leave data compressed if outliers are present within the group (Patro & Sahu, 2015).

6.2.3 Feature Selection and Analysis

Three main methods were used for the feature analysis, the mRMR method, Pearson's correlation coefficient and the F-statistic. The Pearson's correlation coefficient is a measure of the linear association between two continuous variables (Sedgwick, 2012). Pearson's correlation coefficient is calculated as the covariance of the two variables, divided by the product of their standard deviations (Sedgwick, 2012). It is not a perfect

method for feature selection as the Pearson's correlation coefficient only measures linear relationships and may not be appropriate for capturing non-linear relationships (Sedgwick, 2012).

The F-statistic is a statistical measure used to determine whether there is a significant difference between the means of two or more groups (Weir & Hill, 2002). In an ANOVA, the f-statistic is calculated by dividing the variance between the groups by the variance within the groups (Weir & Hill, 2002). In short, the larger the f-statistic, the more likely it is that there is a significant difference between the two groups (Weir & Hill, 2002).

Visualisations were created to illustrate the changes in feature power and the feature correlation when selected with differing methods. Illustrative plots (Figures 7.1-7.7) were calculated to show the correlation between all features and how the use of mRMR method alters the selected features (Zhao et al., 2019a) . Correlation plots were created to show the change in correlation between the top 15 features selected with the mRMR method and when selected with F-statistic alone (Weir & Hill, 2002). A third set of plots was created to show the F-statistic of the top 15 features when selected using the f-statistic alone and mRMR method. To calculate the feature correlation, the Pearson's correlation coefficients were calculated between each of the features and presented within a heat map (Sedgwick, 2012). F-statistic was calculated using the Scikit-Learn package (Pedregosa et al., 2011). This process was repeated for the three main classification tasks within this thesis, namely the true event detection, action recognition, and impact type prediction.

6.2.4 Train-Test Split

For all datasets the data was split such that the training set contained approximately 80% of the data, whereas the test set would contain 20% of the data. For the genuine or spurious classification challenge, this split was achieved using five of the six matches (~80%), and data from one game was used as the testing group. The HAE causation data was split so that four of the matches were used within the dataset, whereas the remaining two games were used within the test set. This meant that no data from the same match appeared in both the training and test groups, although data from the same players may occur within the train and test sets. Feature analysis was only conducted upon the training data to preserve the integrity of the test data. The

purpose of this being that the testing data are to remain as an impartial measure of classifier performance, and to act as a guarantee that trends occurring within the training data are genuine and apparent in the whole population.

When applying the chosen normalisation method to the test data, it is important to note that information must not leak from the test set to the train. Therefore, when scaling the testing data, it must be split before a normalisation algorithm is trained. Once the algorithm is trained on the training data, the test data may be transformed using the trained algorithm.

6.3 Results

6.3.1 Genuine Impact Detection

Initially, the feature set was 100 features selected using the mRMR method. This included features from each category and kinematic measurement, over a wide range of frequencies. This consisted of 20 pulse parameters predominantly measuring the rotational acceleration and velocity, nine higher frequency PSD measures of linear acceleration, and eight derivatives of rotational acceleration and velocity. The remaining features consist of wavelet transformation from all kinematic measurements across the whole spectrum of frequencies used.

Within Figure 6-1 to Figure 6-3 is illustrated the effect of the mRMR method on the correlation and f-statistic. In Figure 6-1 it shows how the features contain high amounts of correlation when in the original groups and how it is redistributed when the mRMR selects features, with the first select in the top left of the figure. Figure 6-2 show the correlation between the top 15 features when selected with f-statistic alone and the mRMR method. The f-statistic showed that nine of the top 15 features were selected which were derived from wavelet transformations. Of these features, z linear acceleration (ZL) the most predictive of the class, with six other wavelet features using frequencies of 60 Hz and below and two higher frequency features included. Pulse parameters were the second most selected group by f-statistic alone, contributing the remaining six features. Pulse width Y-axis linear acceleration (YA) was the only rotational acceleration derived metric and was reported to have the highest f-statistic of all the features, shown in sub-figure E. The addition of the mRMR method to the feature selection process had the effect of reducing the correlation present in the 15 first selected features. It also meant a change of five features selected within this

group, however except the inclusion of one PSD measure, they changed to similar features. Whilst the correlation was reduced, so was the f-statistic of the selected features, illustrated in Figure 6-3.

Figure 6-1: The effect of mRMR on feature selection for the impact detection task. Upper: Correlation between all features. Lower: correlation after all features organised by mRMR.

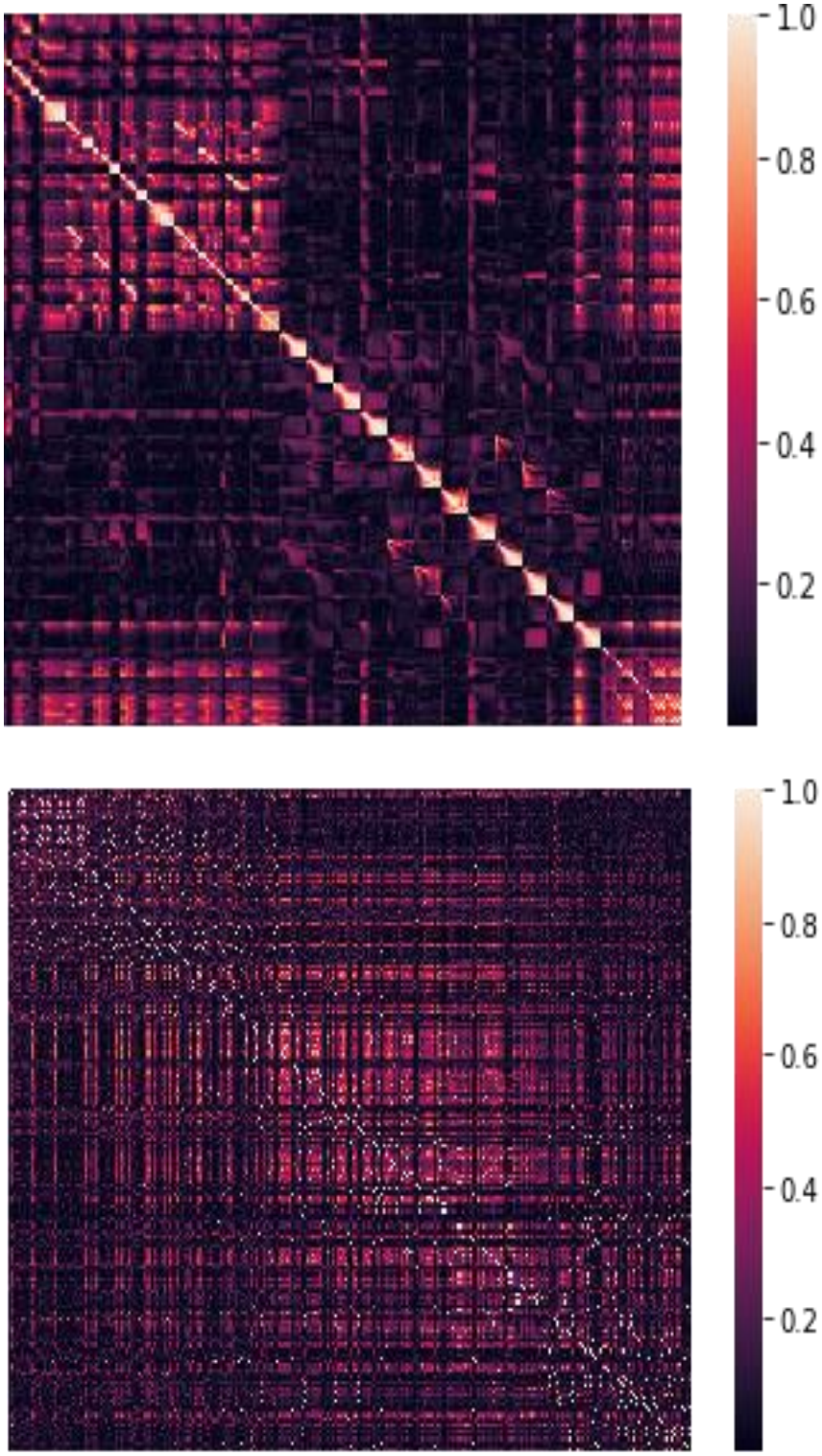


Figure 6-2: Upper - Correlation between top features selected with F-statistic. Lower - Correlation between top features selected with mRMR.

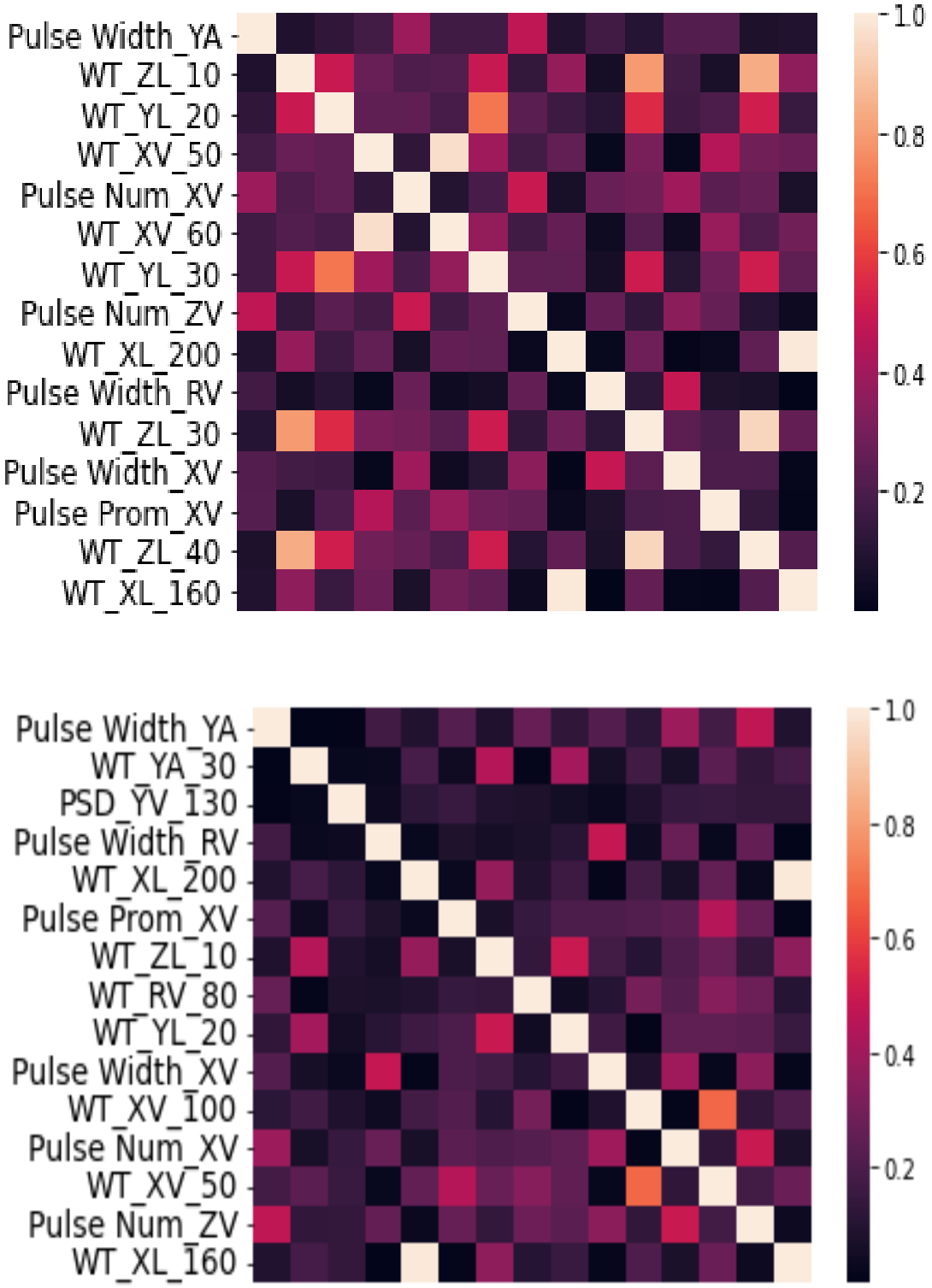
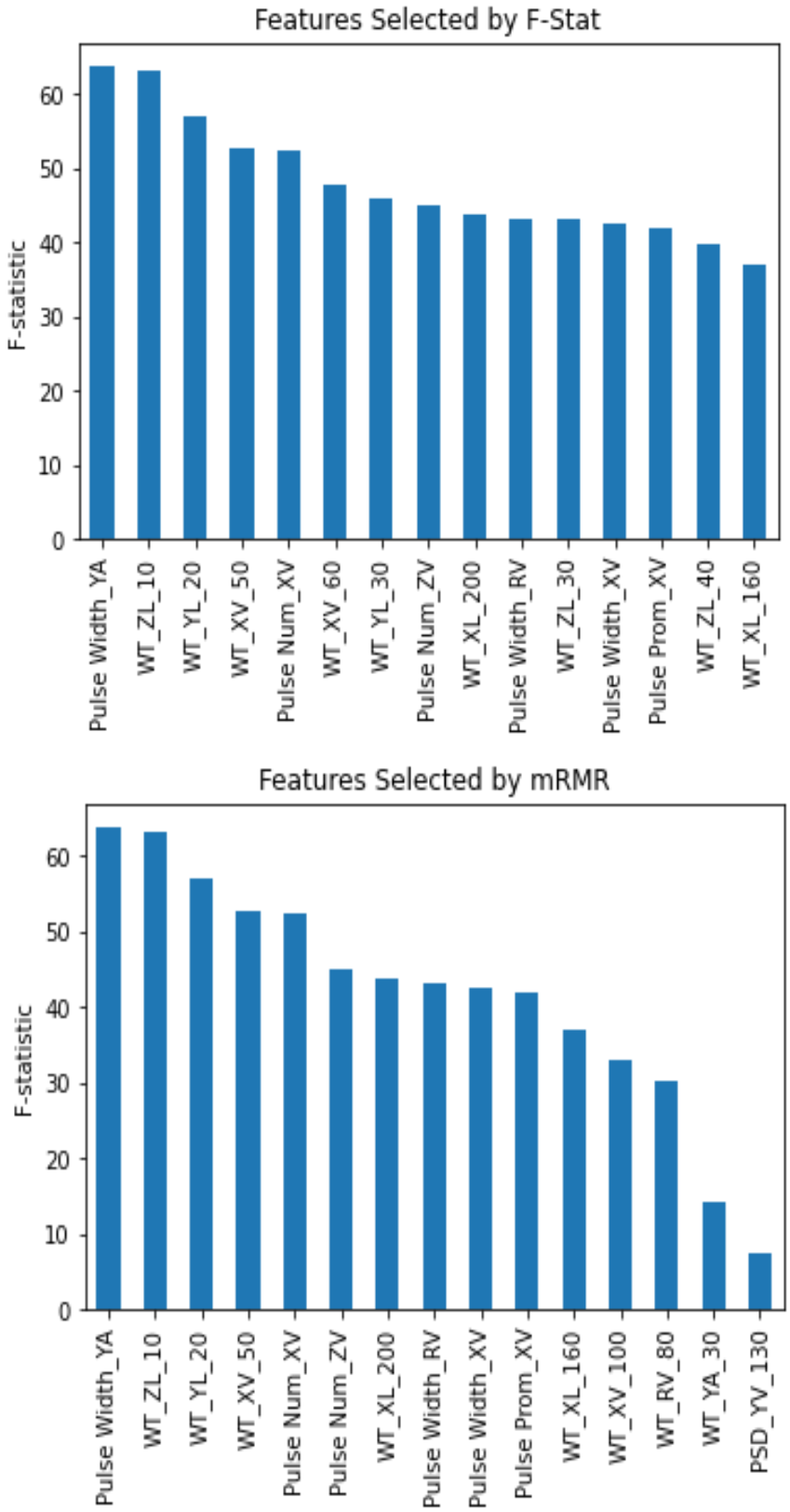


Figure 6-3: Upper - F-statistic of features with highest F-statistic. Lower - F-statistic of top features selected by mRMR.



6.3.2 Action Recognition

The same process was conducted for the two datasets used within the causation analysis tasks, as an effort to create a better understanding of the features used. The top 15 features as selected by the mRMR method are shown in Figure 6-4. These features included a wide range of the generated feature groups, with the frequencies ranging across the whole spectrum of those available (10-200 Hz). Features were derived from each of the types of motion, rotational velocity, rotational acceleration and linear acceleration, and recordings from each axis (X, Y, Z, and R) featured within the data in some form. In terms of feature groups, for the mRMR selected group, six of the features selected were wavelet transformations, three PSD measures with the remaining consisting of statistical measures and pulse parameters.

When selected using only the f-statistic, the chosen features showed a much greater level of correlation than the features selected within the first classification task (impact detection). Features describing similar aspects of the recordings appeared numerous times, for example the inclusion of PSD features at the same or similar frequencies. Only six features appeared within both sets created, shown in Figure 6-5, with the others removed and replaced by features with lower f-statistics. This also corresponds with a greater drop in f-statistic in the mRMR selected group compared to the f-statistic group.

Figure 6-4: The effect of mRMR on feature selection for the action recognition task. Upper: the correlation between features when selected by f-score only. Lower: correlation between features when selected by mRMR.

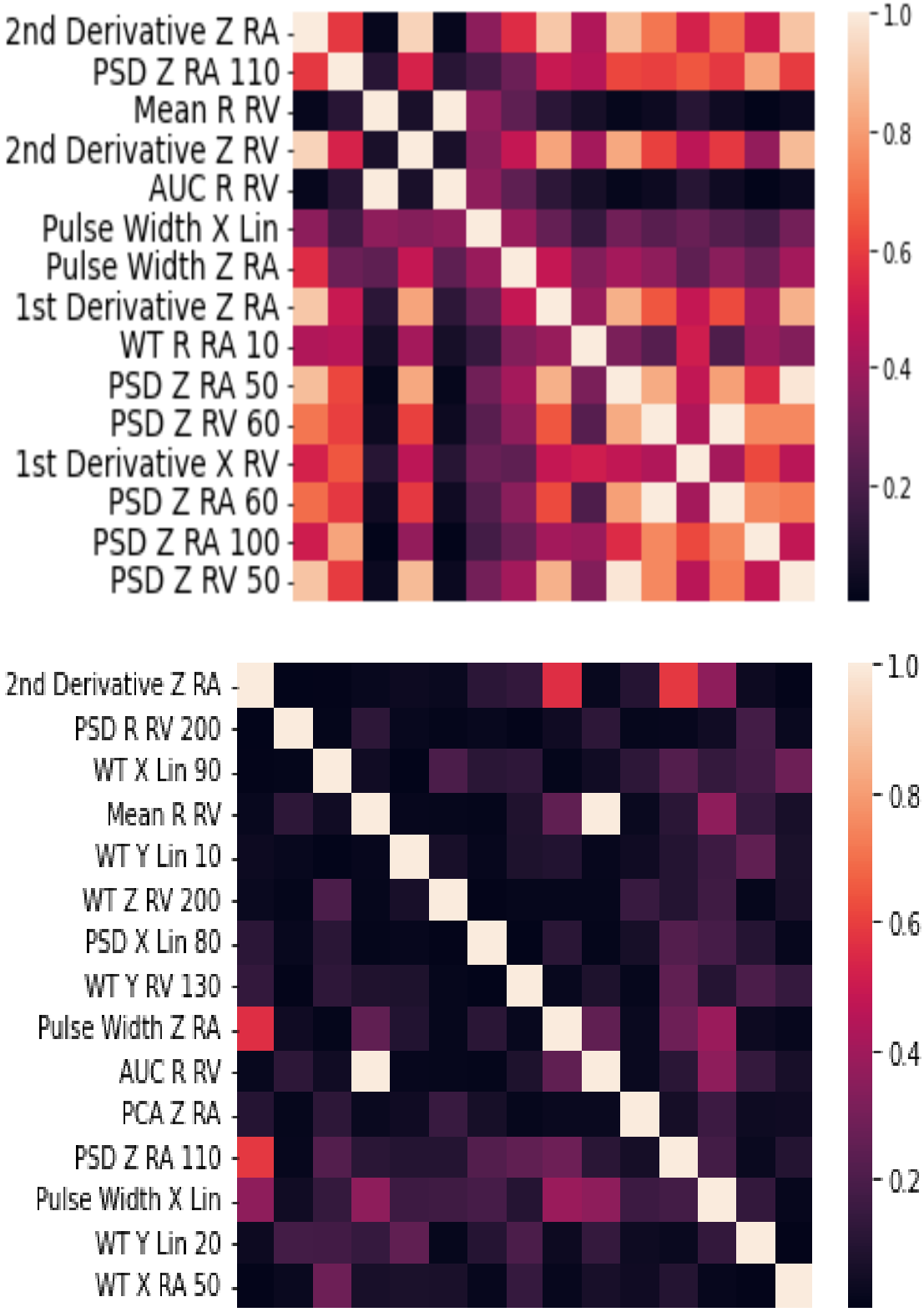
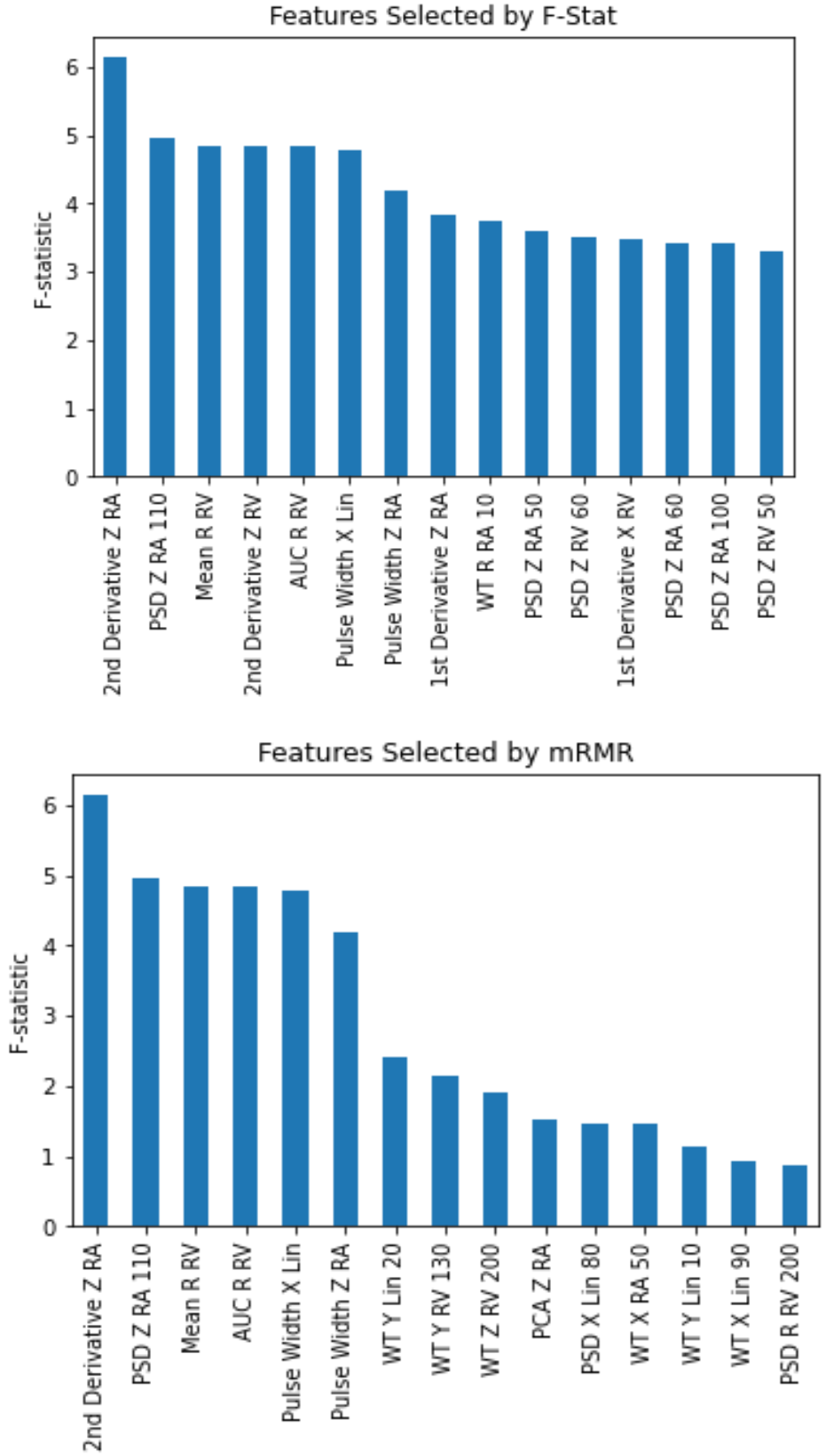


Figure 6-5: The effect of mRMR on feature selection for the action recognition task. Upper - F-statistic of features with highest F-statistic. Lower - F-statistic of top features selected by mRMR.



6.3.3 Impact Type Prediction

The highest performing features in terms of f-statistic and as identified by mRMR are shown within Figure 6-6. For the top performing features, wavelet transformations made up 5 of the top 6 performing features and 7 of the top 15. Three PSD measures were also apparent, with the remaining made of statistical measures and pulse parameters. As with the previous classification task, the frequencies used within the features ranged across the full spectrum of available features, and features were derived from linear acceleration and rotational acceleration / velocity. Six features were retained in both the mRMR selected group and F-statistic derived group, with the features included by the mRMR method having a lower average f-statistic (Figure 6-7). However, Figure 6-6 shows high levels of correlation between the top-performing features within this task, with six of the highest performing features derived from the wavelet transformation of the linear acceleration, with the features including the range 160-200 Hz in addition to the 10 Hz transformation. After the use of mRMR, these were reduced to two features, using 10 and 200 Hz.

Figure 6-6: The effect of mRMR on feature selection for the Impact type task. Upper: the correlation between features when selected by f-score only. Lower: correlation between features when selected by mRMR.

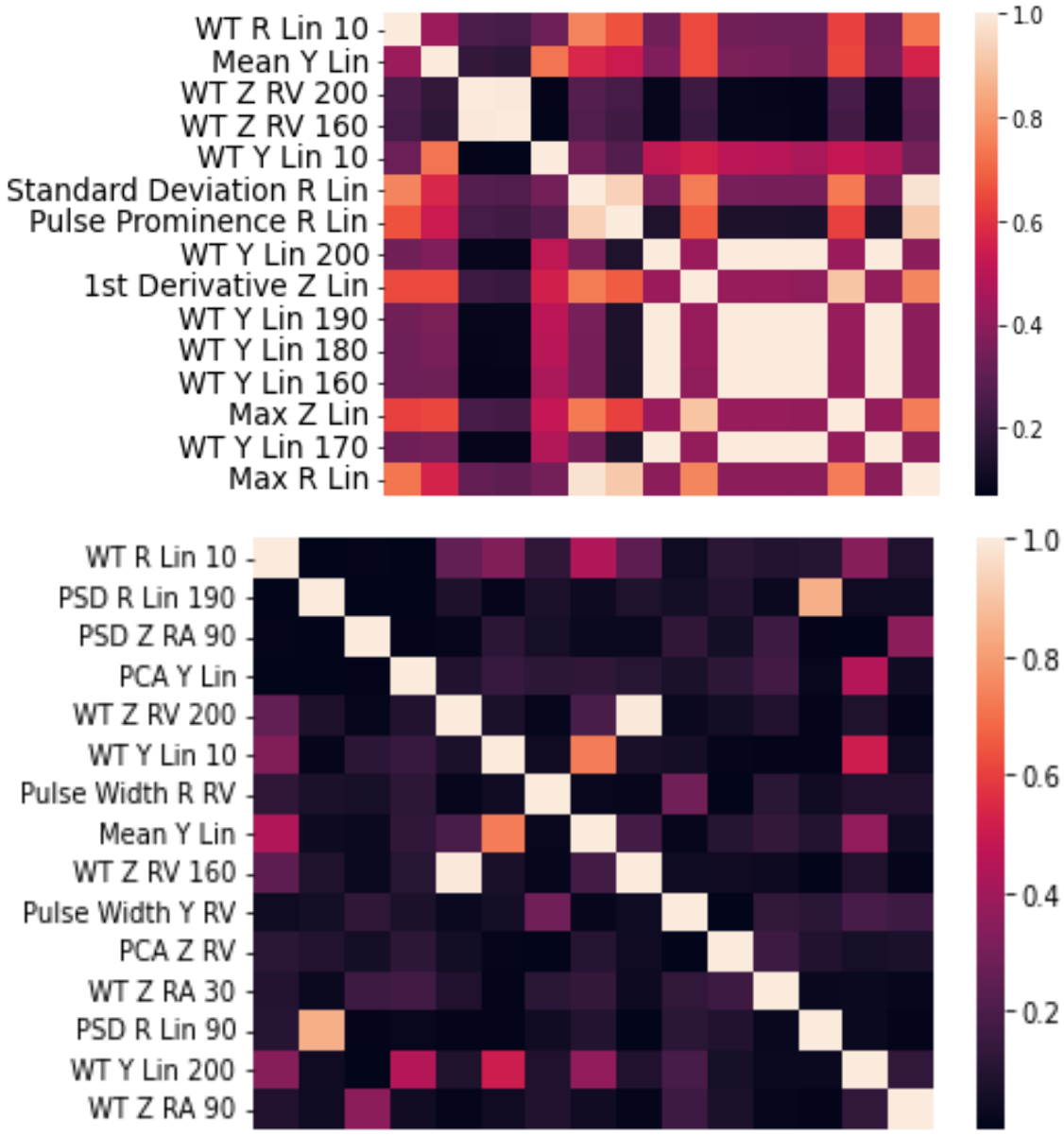
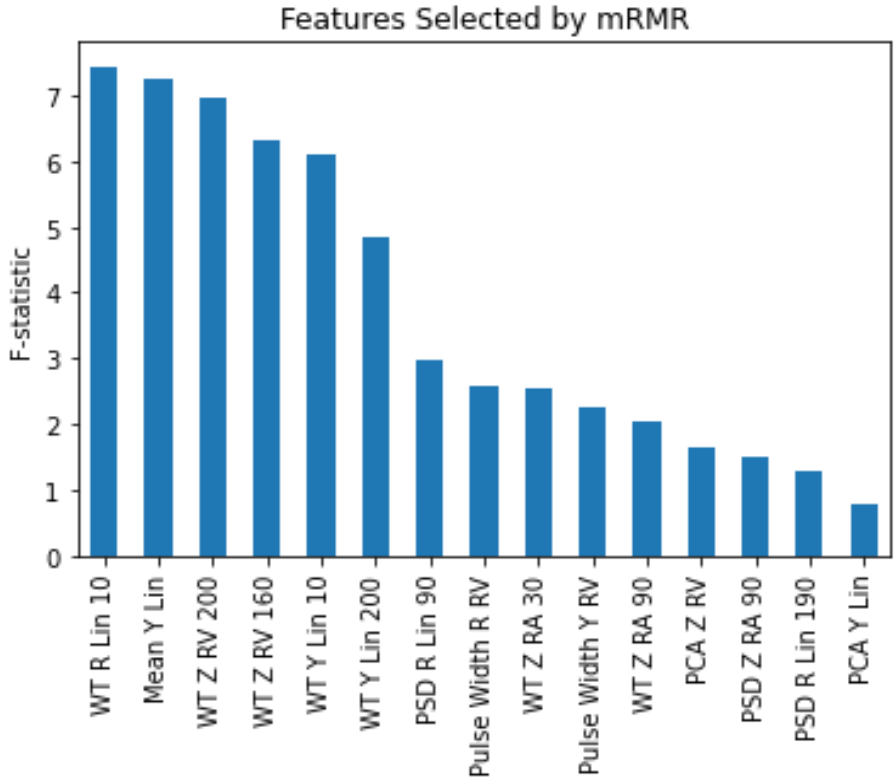
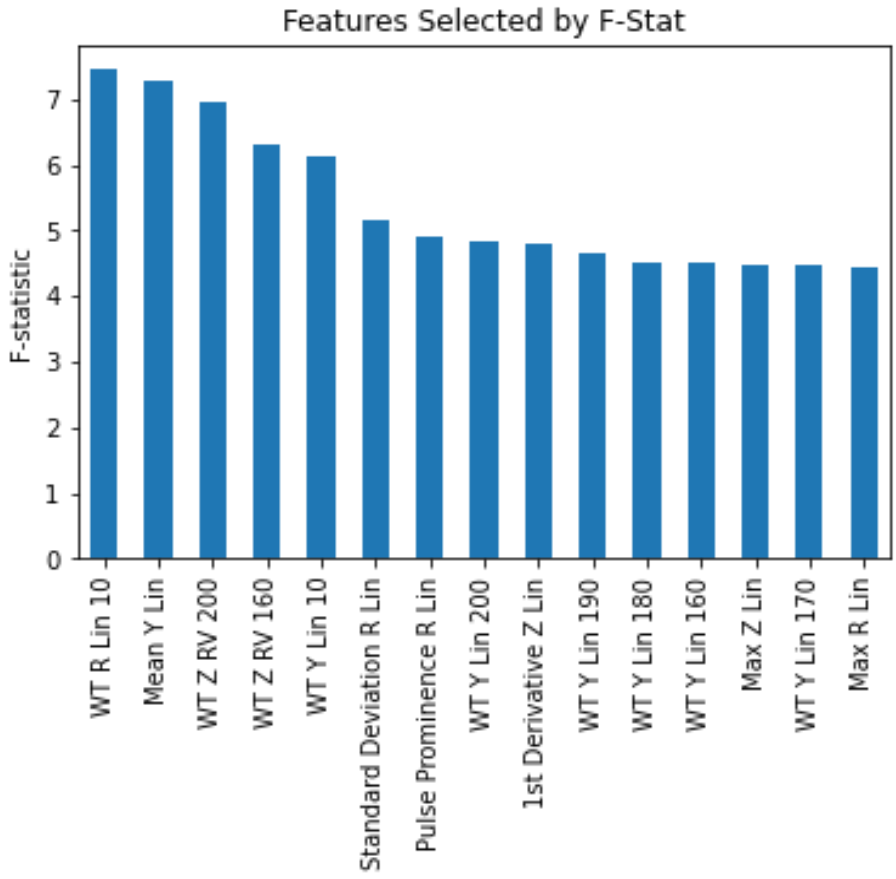


Figure 6-7: The effect of mRMR on feature selection for the Impact type task. Upper: F-statistic of features with highest F-statistic. Lower: F-statistic of top features selected by mRMR.





6.4 Discussion

The primary purpose of this section was to transform the data into a robust set of features for each classification tasks. These classification tasks were the detection of genuine impacts, the detection of direct contact to the head, and the prediction of the action the player was conducting when the head acceleration occurred. To develop an effective feature set, the mRMR method was used to select predictive features whilst picking features with lower correlation to train the classifiers. Additionally, this aspect of the study aimed to not only to develop effective features, but to start developing an understanding of the features were valuable for classification tasks within this domain.

When creating ML classifiers with smaller datasets, as in the following studies, efforts must be made to control over-training. Models created with high numbers of features and low counts of data points may learn spurious patterns within the dataset that are not present in reality (Hua et al., 2005). Recommendations have been made that a feature count of square root of the number of data points (N) to N is suitable. This depends on the level of correlation within the feature set and on the assumption all of the features are informative to the task (Hua et al., 2005). Therefore, the optimal value of features will be dependent on the number of data points and the quality of the features (Zheng & Casari, 2018).

An important consideration in feature selection is minimizing correlation among features since highly correlated features often contain redundant or similar information. This redundancy not only fails to contribute new valuable information to the model but can also increase computational load and the risk of overfitting. Overfitting occurs when the model learns noise and specific patterns from the training data that do not generalise well to new, unseen data. Moreover, the presence of highly correlated features can introduce model instability, where minor changes in the data or model parameters result in significantly different model outcomes. This instability can lead to a less robust model, affecting its reliability when applied to new data. Additionally, correlated features can diminish the interpretability of the model by making it challenging to discern the distinct impacts of individual features on predictions, thus reducing the clarity of model explanations.

This was particularly prevalent within the PSD and WT features were created sequentially across frequency spectrums, capturing similar information. The narrow frequency bands led to the leakage of information between these frequency bands. This suggests performance may have been improved by reducing the number of frequency bins. This would have the benefit of improving interpretability and should other studies follow the same approach will allow for the direct comparison of feature values between studies which may aid in future research. This approach may have the negative effect of reducing the performance of classifiers as valuable information may be lost.

An alternative would be the use of PCA to reduce the dimensionality of these feature groups, although this has its drawbacks. Once features have been transformed using PCA, the level of abstraction is increased and therefore the features are less interpretable. An alternate method that may have been used is mutual information. Mutual information is a measure of the statistical dependency between two variables and can capture dependencies beyond linear relationships, including nonlinear associations. However, it does not explicitly deal with or account for multi-collinearity or correlation between features in the same way as traditional correlation metrics like Pearson's correlation coefficient. This is perhaps more computationally expensive and has not been shown to perform significantly better than the FCQ variant of mRMR (Zhao et al., 2019a).

The FCQ variant was selected to reduce the features as it was described by the creators as a suitable method for most feature sets. This method uses a combined metric based on the f-statistic and Pearson correlation coefficient to select features. This objective could be achieved through the use of f-statistic alone; however, this could lead to a feature set that contains groups of highly correlated features. Highly correlated features are rarely valuable to classification tasks in ML and can be damaging to performance. This could be particularly problematic in this study, as many similar features are created as they were created incrementally leading to this correlation. Therefore if features are reduced to a particular number, then this may lead to redundancy amongst the features.

6.4.1 True Positives

When detecting true positive head accelerations, mRMR feature analysis identified pulse parameters and wavelet transformations as the most valuable feature categories, providing both the greatest number, and most powerful features for classification. The frequencies used to train the classifiers ranged from 10-200 Hz, with features closer to the 10 and 200 Hz limits typically appearing earlier in the order of selection. The appearance of high-frequency features in this classification task is in agreement with another publication that linked high-frequency signals with unreliable recordings (Luke et al., 2024). This is in agreement with studies conducted with an SVM classifier to detect genuine events. These studies showed that features with low-frequency values (10-30 Hz), were predictive of genuine events, perhaps due to these frequencies' proximity to voluntary human motion frequencies (Khusainov et al., 2013; Wu et al., 2014, 2018). However, low frequencies have not always been shown to be predictive of genuine HAEs. Goodin et al., 2021 reported SHAP values that indicated that high PSD values for rotational acceleration in the frequency bands of 30-50 Hz indicated spurious events, while 1-40 Hz linear acceleration indicated genuine events. The reasoning for this could be device-specific consideration about how vibration is transferred to the kinematic sensors.

Further investigation into this topic would be beneficial from a mouthguard design perspective. A greater understanding of the mechanisms that cause spurious events will lead to better designs of hardware, filters, and algorithms to ensure that future devices continue to improve their reliability.

6.4.2 Causation analysis

From the action recognition study, wavelet transformations were identified by the f-statistic and mRMR as the most predictive feature categories. This included features derived from each type of motion and each orientation. However, a diverse range of other features were included, both those typically associated with impact verification and the features derived from specific action recognition studies. It was shown that frequencies closer to the feature extremities (10 and 200 Hz) appeared more often than those in the middle of the range. It was seen that there was a sharp decline in feature performance across the highest performing features identified. The f-statistic showed a drop off after the top five to six features, indicating that few are suitable for this task. As with many tasks where the statistical significance of the difference between groups

is low, then more data should be collected to find more significant unique features before rebuilding the classifier to generate better results. Should increasing the data not prove sufficient it may prove that the classification task is too challenging with the current sensor designs and they may require altering should this classification task be achieved.

7 Development of a Head Acceleration Event Classification Algorithm for Female Ru

The majority of this chapter was published within the following paper: Powell, D. R. L., Petrie, F. J., Docherty, P. D., Arora, H., & Williams, E. M. P. (2023). Development of a Head Acceleration Event Classification Algorithm for Female Ru. Annals of Biomedical Engineering. https://doi.org/10.1007/s10439-023-03138-9. David Powell was the lead investigator, supervised by Dr Arora, Dr Williams, and Professor Docherty. Freja Petrie assisted with data collection.

7.1 Introduction

Video verification is a time-consuming activity with the quality of the output being dependent on the skill of the reviewers and the quality of the footage used (Cortes et al., 2017; Patton, Huber, McDonald, et al., 2020). Furthermore, ML algorithms are reliant on high-quality datasets, which in turn rely on both video verification and the quality data recorded by the HIT device. Additionally, questions remain over the on-field performance of algorithms trained on laboratory data (Kieffer et al., 2020; Patton, Huber, McDonald, et al., 2020). Despite this, it has been reported that only 36% of head impact studies use video verification, whilst 74% use filtering algorithms (Patton, Huber, Jain, et al., 2020). Multiple studies have proposed ML algorithms to improve the on-field performance of IMGs in specific sports, achieving near-human-level performance (Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2014, 2018). This illustrates that for ML algorithms to become a viable injury detection method, they must be derived from data collected from appropriate populations.

One area where the population differences may be particularly prevalent is between males and females, due to the differences in the cervical spine and the kinematic response of the head during impact (Caccese et al., 2018; Mohan & Huynh, 2019; Stemper et al., 2009; E. M. P. Williams et al., 2021). The structure of the male cervical spine results in greater stability and resistance to external loading than the female cervical spine (Mohan & Huynh, 2019). This results in females experiencing an increased magnitude of displacement and acceleration during vehicle collisions and sporting contact events (Caccese et al., 2018; Stemper et al., 2009; E. M. P. Williams

et al., 2021). As the motion of the head is reported to be different during different during these events, it is reasonable to assume that the trends developed by ML algorithms may be sex specific.

Despite the rapid growth of female RU participation and the increasing number of professional players, no algorithms for female sport or RU have been specifically developed (Woodhouse et al., 2022; World Rugby, 2019). Therefore, it is imperative that this gap is addressed to ensure player safety and research quality (E. M. P. Williams et al., 2021). This study aims to develop a classification algorithm to detect head accelerations from IMGs, in female RU players.

7.2 Materials & Methods

7.1.1 Data Collection

This study uses data collected and processed in Chapters 5 and 6, where full descriptions of the processes involved are reported.

7.1.2 Classifiers and Features

Four classification algorithms were trained to determine whether the filtered accelerations were impact events or false impact events. Due to the large number of classification algorithms available, the classifiers selected for this task had previously shown success in head acceleration classification tasks. These algorithms were an Adaboost decision tree (3.4.6), support vector machine (3.4.4), and two extreme gradient boosted decision tree models (3.4.6), CatBoost and XGBoost, as used by Wu et al., (2018), Gabler et al., (2020) and Goodin et al., (2021).

The classifier determined key patterns in the descriptive features (433.3.1) of the filtered six-axis kinematic data. Features were grouped into four categories, pulse parameters, positional derivatives, PSD, and wavelet transformations. Except for positional derivatives, the feature categories have been used previously to train HAE classifiers (Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2018). Analysis was undertaken in a Python 3.8.10 computational framework (3.1).

The prominence, width, and number of pulses were identified from local maxima in the signal (3.3.1). The prominence of each peak was measured by calculating the vertical distance between the highest point and the lowest contour line. The width of each peak was measured by calculating the horizontal distance at the lowest contour

line. In the event of there being multiple instances of either measure, the maximum value calculated would be used. The final measure was the total number of peaks per signal.

The first and second derivatives were calculated from each of the kinematic measurements. The first derivative was calculated from the change in two sequential recorded values, with the second derivative calculated by the same process as the first derivative. The maximum absolute value of each signal used, this provided the maximum rotational acceleration and jerk as calculated from the rotational velocity, and the respective jerk and snap as calculated from the linear and rotational accelerations.

The PSD (3.3.1) describes the power of a signal in frequency components. This was calculated in 20 Hz windows using Welch's method between 20 and 200 Hz, with the upper bound determined due to the filter's cut-off frequency. The power values for each frequency were used as a feature, providing ten features per vector.

A wavelet transformation (3.3.1) was used to provide time dependant frequency analysis. A CWT was conducted for each signal using the Ricker wavelet function between 10 and 200 Hz, in 10 Hz increments. The strength of frequencies calculated at the recording's maxima were used as features.

The features were appended to a data frame, in which they were scaled to have zero mean and a standard deviation of one. To reduce overfitting and training times, the total number of features was reduced to 100 using the FCQ variant of the mRMR method (Zhao et al., 2019a)(3.3.2). There are no standardised rules on the appropriate number of features that should be considered. In this study, a value closer to the lower limit of features outlined Hua et al. 2005 was selected (Hua et al., 2005). This was to combat the prevalence of collinearity between the features of the dataset (Hua et al., 2005).

7.1.3 Performance Metrics

The AUROC and AUPRC were used as the performance measures of the models. Area-based metrics measure the classifiers' ability to discriminate between events at all decision thresholds, which in turn gives a measure of how well the classifier has separated the data (Mitchell, 1997). This provides greater detail of the classifier's performance than measures that require a specific prediction threshold. More details

of these metrics can be found in section 3.5.2. Area based performance metrics have been previously identified as an important metric in HIT due to its value when evaluating imbalanced datasets (Saito & Rehmsmeier, 2015; Wu et al., 2014, 2018).

SHAP values were calculated to analyse the effect of feature values on the classifier's outputs (3.5.1) (Lundberg & Lee, 2017). SHAP values are calculated from the prediction of labels on feature vectors constructed from feature values and combinations found within the dataset. The effect of the feature value can then be measured along with its effect on label prediction.

7.1.4 Classifier Selection and Development

The RU data was roughly split into a training group using five of the six matches (~80%), and data from one game was used as the testing group. No data from the same session appeared in both the training and test groups.

The first training stage was the simultaneous hyper-parameter and feature selection, conducted using an eight-fold cross-validation. During cross-validation, the training data was split into eight approximately equal groups, with all but one of the groups used to train the classifier and the remaining groups used to assess the classifier's performance. The cross-validation process works so that all data appears in the validation group once. A dictionary of hyper-parameters was created, and an exhaustive search of all combinations was used to find the highest performing classifier hyper-parameters, with models optimised for AUROC. This was initially tested with ten features then re-run with ten features added, until the feature set had reached the maximum size of 100 features. Feature optimisation was conducted in this manner to reduce the likelihood of overtraining the classifiers, whilst optimising the features and hyper-parameters. The hyper-parameters of the highest performing models were then recorded. These configurations were then retrained on the entire training dataset ten times, with the highest performing model then used to predict labels on the test data. AUROC, AUPRC and SHAP values were then calculated to assess and explain model performance. The code and trained classifiers may be made available following a reasonable request to the corresponding author.

7.3 Results

A total of 214 HAEs and 466 spurious events were identified from the raw data during the video verification process. The training data was formed of five matches,

containing 166 head accelerations events and 400 spurious events, with the remainders used for the test data.

The 100 features selected using the mRMR method included features from each category and kinematic measurement, over a wide range of frequencies. This consisted of 20 pulse parameters predominantly measuring the rotational acceleration and velocity, nine higher frequency PSD measures of linear acceleration, and eight derivatives of rotational acceleration and velocity. The remaining features consist of wavelet transformation from all kinematic measurements across the whole spectrum of frequencies used. A summary of the most informative 25 features obtained is shown in Table 7-1.

Table 7-1: Highest performing features and their orientation on the head impact classifier models.

Feature Category	Recording	Direction and Frequency			
		X	Y	Z	R
Wavelet Transformations	Linear Acceleration	200, 160	20, 30	10, 50, 30, 40	
	Rotational Velocity	100, 50, 60, 80	160		100, 80
	Rotational Acceleration		30		
Positional Derivatives	Rotational Velocity	PP, PN, PW	PW	PN, PW	PW, PN
	Rotational Acceleration	PP	PW		

The best performing 20 features for the CatBoost and SVM classifiers as identified by SHAP values were found. The most important features for classification were predominantly pulse parameters and wavelet transformations at very low or high frequencies. For the CatBoost classifier, wavelet transformations made up over half of the 20 features used. The transformations in the X direction at 20 Hz and Y direction at 200 Hz for the linear classifier were the most important features. These important features and their relative contributions to classification of head impacts are provided in Figure 7-1. SHAP values were also calculated for the SVM classifier, with the top 20 features primarily consisted of pulse parameters, with the remainder of features being wavelet transformations, the results are shown in Figure 7-1 and Figure 7-2.

Figure 7-1 – The 20 most valuable features identified through SHAP for classification for the CatBoost classifier.

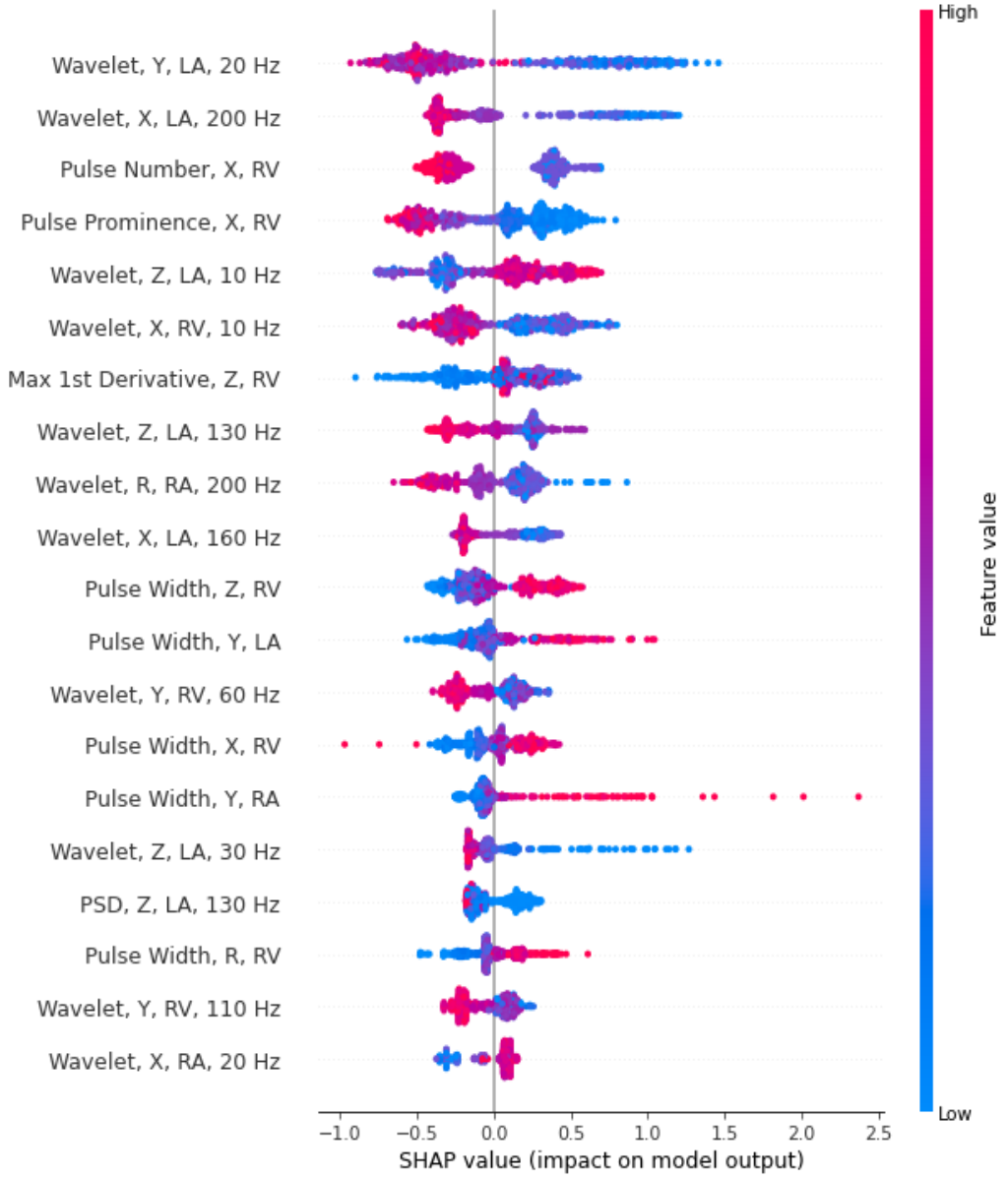
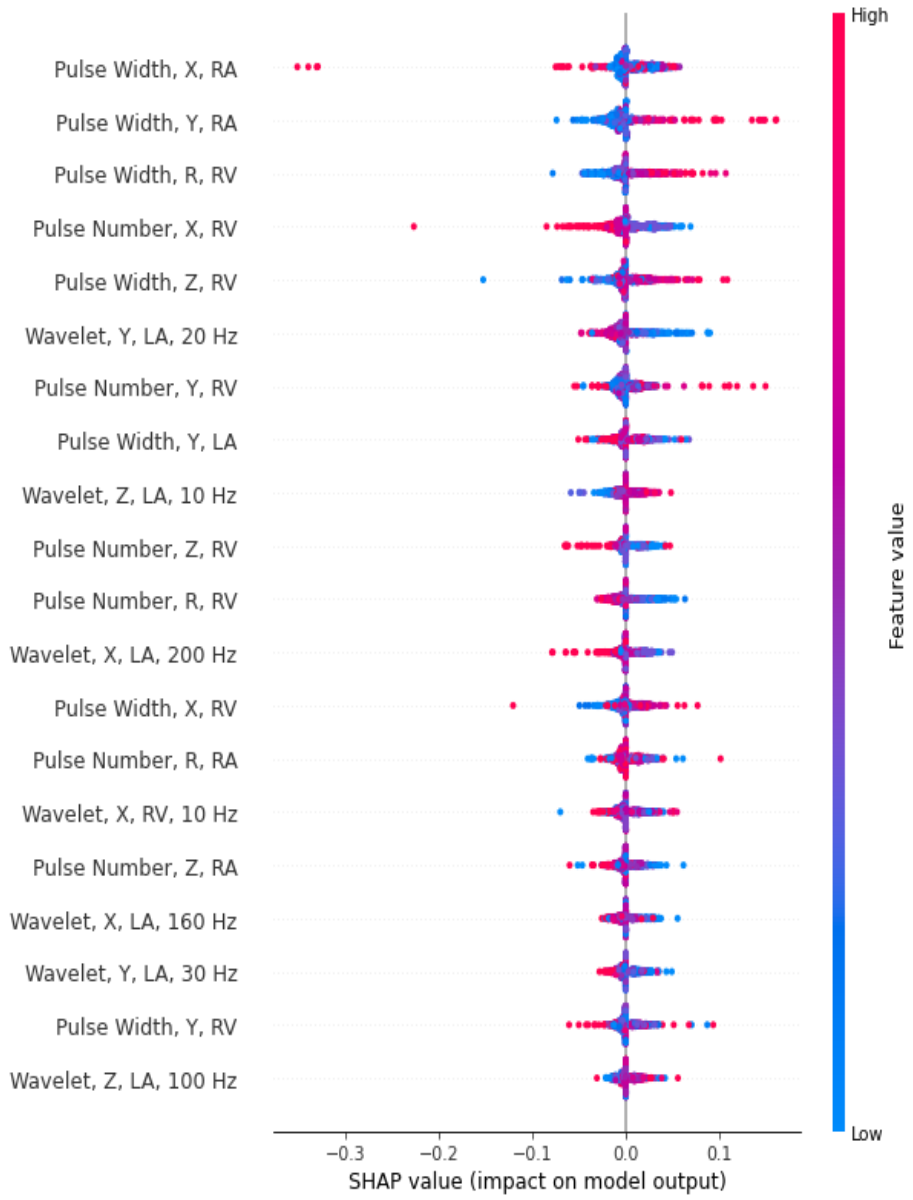


Figure 7-2: The 20 most valuable features identified through SHAP for classification for the support vector machine classifier.



During the optimisation process, the CatBoost classifier achieved the highest cross-validation AUROC score of 0.900, with the next highest performing classifiers being XGBoost, SVM and Adaboost decision tree respectively. When tested upon the test dataset, the SVM classifier achieved the highest AUROC and AUPRC, followed by the CatBoost, XGBoost and Adaboost DT. The results of all the tests are shown in

Table 7-2, with all receiver operator curves and precision recall curves plotted in Figure 7-3 and

Figure 7-4.

Table 7-2: Classification performance of various models in cross-validation and validation within the test dataset

Model	Cross-Validation optimisation Score (AUROC)	Test Score AUROC	Test Score AUPRC
SVM (70 Features)	0.885	0.92	0.85
Adaboost DT (100 Features)	0.866	0.89	0.81
XGBoost (20 Features)	0.892	0.89	0.82
CatBoost (90 Features)	0.900	0.91	0.82

Figure 7-3: Classifier precision recall curves. This represents the trade-off between capturing exclusively genuine events (precision) and capturing all genuine events (recall). This is computed by calculating the precision and recall as the model's decision threshold varies.

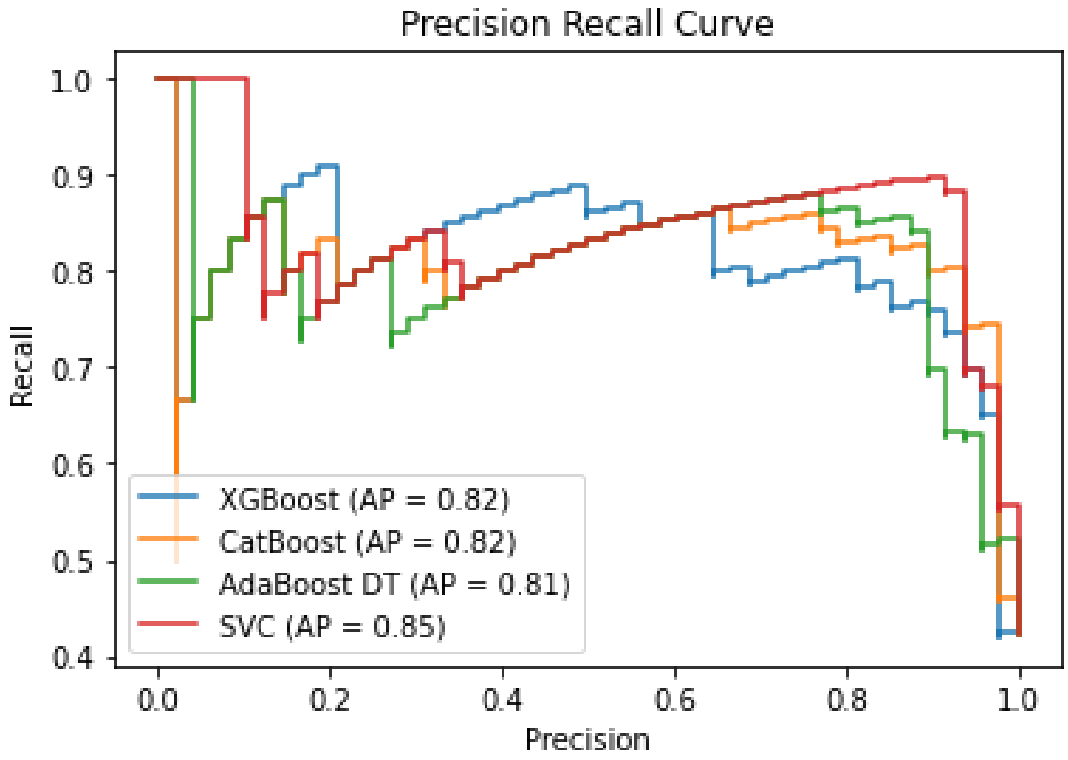
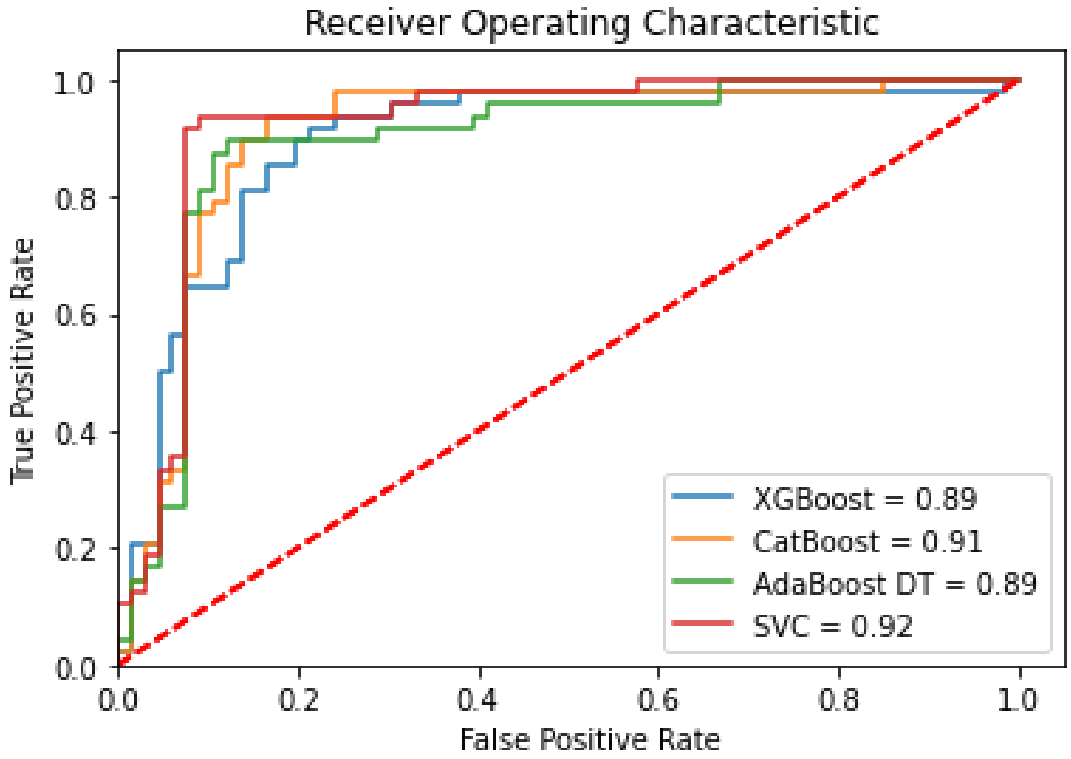


Figure 7-4: Classifier receiver operator curves. This represents the trade-off between capturing exclusively genuine events (precision/true positive rate) and excluding false positives (false positive rate). This is computed by calculating the true and false positive rate as the model's decision threshold varies.



7.4 Discussion

In this study the performance of four classification algorithms trained on HIT data from women's collegiate RU were evaluated. This work developed a classification algorithm to detect head accelerations recorded by IMGs, in female RU players. In total, four head acceleration classification models were developed. This is the first study to both develop classifiers with exclusively female data, and to classify impacts from RU. Each model performed well in the classification task, with a SVM algorithm providing the greatest performance when tested, with an AUROC and AUPRC of 0.92 and 0.85 respectively. This study represents an important step in the development of female specific head acceleration detection algorithms, which will contribute to safer participation in RU and more reliable study of female contact sport.

SHAP and mRMR feature analysis identified pulse parameters and wavelet transformations as the most valuable feature categories, providing both the greatest number, and most powerful features for classification (Figure 7-1, Figure 7-2). The frequencies used to train the classifiers ranged from 10-200 Hz, with features closer

to the 10 and 200 Hz limits typically appearing earlier in the order of selection. Both high and low frequencies contributed to predictions of the negative classes (Table 7-1). High feature values of wavelet transformations that had characteristic frequencies at the feature limits generally led to negative class label prediction. Conversely, features that led to positive prediction were limited to those based on low frequencies and some pulse parameters. For example, 10 Hz motion in the vertical direction (z) was predictive of actual events, as was pulse width in all directions (Figure 7-1).

This analysis highlighted the importance of the feature's direction and the characteristic frequency. A feature with characteristic frequency of 20 Hz in the y-direction, for example, was strongly negative in its contribution to prediction of actual impacts, whereas a feature with characteristic frequency of 10 Hz in the z direction was strongly positive in its prediction of actual impact (Figure 7-1). This may occur as the low frequency features may be capturing voluntary human motion, which generally occurs at frequencies below 10 Hz (Khusainov et al., 2013). For example, spurious events can occur while players are at rest and motion largely restricted to the x and y axis. As a player runs or enters contact, there will be a greater movement in the z direction, which is registered more often in positive events, hence leading to a positive label prediction.

This study used a novel data set collected from 25 IMGs from adult female rugby players in a university, 1st team squad. This data was collected in accordance with the methodology outlined by Williams et al., (2021). The last six games of the season were monitored, with the number of players in the match day squad possessing IMGs ranging from 12 to 22 per game. In total, over 7,500 player minutes of data was gathered, with every field position represented. Hence, the data is broadly representative of female rugby at the penultimate level. Each game was verified independently by two reviewers, who while blinded to each-other's classifications, achieved a high level of agreement. In total 680 impacts were used within the training and testing of the classification algorithms, which is in line with the datasets used in the studies of Gabler et al., (2020) and Wu et al., (2018). The Prevent Biometrics hybrid IMGs performed comparably to the on-field performance tests reported by Kieffer et al., (2020).The algorithms saw consistent performance across the training and test datasets reaching values proximal with those in the literature, indicating a reliable study.

The classification results shown in Table 7-2 are, by some metrics, slightly lower than some studies of American football and Australian rules football head accelerations (Domel et al., 2021; Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2018). Specifically, Goodin et al., (2021) recorded maximum recall of 94.7% and 95.7% for genuine and spurious events in Australian rules football. Whilst the American football classifiers of Domel et al., (2021), Gabler et al., (2020) and Wu et al., (2018), achieved precisions of 98.3%, 93.8% and 86.0%, with recall values of 87.2%, 100% and 76% respectively. For comparison, in this study a peak precision of 89.9% and recall of 91.1% were achieved with the SVM classifier. Note that the comparisons across sports lack equipoise to make summative assessments of the algorithms considered. In particular, classification of head accelerations in RU requires capture of a variety of head acceleration types. Hence, such models need to be broad, with many features to capture the various head acceleration types. This model broadness provides more potential for spurious data to correlate with one of the features, and lead to a false positive. It is reasonable to assume that sports with a limited range of head acceleration types can classify events with fewer features and limit the potential for false correlation.

The complexity of head accelerations in RU may also manifest as classification errors during the video verification process, as matches were typically filmed without redundancy and from the touchline. While some events were excluded due to being unclear in the match footage, some events may have been misconstrued as the incorrect label by the video verifiers. Additionally, some phases of play such as tackling, rucking, and mauling occur can lead to multiple recordings during the event, which in turn makes classification difficult. In these events, the recording where no clear movement of the head could be seen on camera were excluded from the study. This resulted in a dataset containing only events recorded when clear head acceleration was seen, and events with no head acceleration. Some previous studies have included only direct head contacts, so perhaps the inclusion of indirect events led to the inclusion of more spurious events or a more difficult classification task (Gabler et al., 2020; Wu et al., 2018).

A further limitation to this study, is the lack of availability of testing data acquired from a separate cohort. Whilst the results were consistent across the cross validation and testing, further validation from new end users would assist in assuring the

generalisability of the developed algorithms to female RU cohorts. Furthermore, care should be taken when extrapolating the results of this analysis beyond the cohort tested. In particular, these results may not be applicable in adolescent female rugby, in the professional levels, or in other women's sports. With the classifiers previously validated in other sports, it appears that decision tree-based ensemble and SVM algorithms are both capable of creating high performing classifiers (Gabler et al., 2020; Goodin et al., 2021; Miller et al., 2018). The mRMR method was used to reduce the feature count from 540 to 100 to aid classifier training. This proved successful as the features identified as most important through mRMR were later confirmed as highly valuable by the SHAP values (Figure 7-1). Figure 7-3 shows the SHAP values of the most valuable feature in the Catboost classifier (Wavelet Transform, Y, Lin Acc, 20 Hz) was far more powerful than the 20th most predictive feature. Hence, the pruning approach that was utilised to mitigate overfitting, was unlikely to lose significant amounts valuable information. This retention of performance is shown in Figure 7-3. Whilst this unbalanced contribution of feature strength wasn't as pronounced in the SVM approach, there was still a reduction in feature power across the top 20. Overall, the distinct SHAP values for each feature across the four algorithms tested highlights the importance of providing the algorithms with diverse array of features for head impact classification tasks.

This study has illustrated that it is possible to create high performing HAE classifiers for female RU. This will aid future researchers to more quickly and accurately identity HAEs within female RU. It has been previously reported that there are sex differences between the male and female cervical spine and sex differences in measured head peak-kinematics during RU matches (Alsalaheen et al., 2019; Mohan & Huynh, 2019; Stemper et al., 2005; E. M. P. Williams et al., 2021). Further to this, there has been no cross-validation study of head acceleration classification algorithms across sports or sexes. Such cross-validation is required to establish the repeatability and generalisability of model-based interpretation of IMG data for head acceleration monitoring. Until this is done, it is reasonable to assume that there may also be differences in impact characteristics and that a "one size fits all" approach to classification is not appropriate. As head acceleration classification algorithms learn specific patterns and trends, the patterns learnt from male sport, may not provide replicable results in female sport. Failing to address this may lead to significant

differences in the understanding and identification of brain injury in female sport. With the rapid growth and progressive professionalisation of female RU, it is essential to create specific state-of-the-art head acceleration classification models to provide reliable data, and to protect players (Woodhouse et al., 2022; World Rugby, 2019).

As the methods to collect genuine events have been described within this chapter, the next chapter moves to the next stage of the pipeline shown in Chapter 4. Chapter 8 describes the methods, results and importance of accurately reporting the linear acceleration magnitudes from genuine HAEs.

8 Optimising Head Acceleration Reporting Locations

8.1 Introduction

IMGs measure the kinematics of an athlete's head during potentially injurious head accelerations using a combination of kinematic sensors (Bartsch et al., 2014; Jones et al., 2022). Multiple studies have validated IMGs against reference systems, demonstrating comparable performance to the reference sensor in laboratory settings (Bartsch et al., 2014; Camarillo et al., 2013; Greybe et al., 2020; Stitt et al., 2021). In order to compare the kinematic recordings of the IMG and reference sensor, data must be rotated to align with the reference sensor and translated to estimate the linear acceleration at the reference sensor. However, these transformations assume a fixed geometric relationship between the sensor and the head’s CG, which may not hold in real-world applications.

Given the natural variation in head sizes and morphologies across individuals, a generic alignment and translation method cannot accurately estimate accelerations at the CG for all subjects. As noted by Bussey et al. (2022), reporting accelerations directly from the sensor is insufficient because it fails to account for the complex dynamics introduced by individual anatomical differences, leading to potential inaccuracies in kinematic data interpretation. Bespoke measurements or individualised calibration procedures may be necessary to ensure accurate alignment and scaling of IMG data to the head's CG.

This chapter will investigate the influence of head size on the accuracy of linear acceleration measurements derived from IMGs, assessing the need for subject-specific adjustments in data processing methodologies to improve kinematic reporting fidelity.

8.2 Methods

To measure the distance between the sensor and the CG of the head, along the X, Y, and Z axes as defied within the SAE J211 documentation (Society of Automotive Engineers, 2003). These measurements were then used to translate the linear acceleration using a rigid body transformation. These measurements were used to translate linear acceleration data from video verified, filtered, HAEs to the estimated CG locations, where the effect of the head size could be analysed.

8.2.1 Initial Feasibility Investigations

Before commencing the study, a preliminary investigation was undertaken to assess certain anthropometric measures intended for collection. The aim was to determine the feasibility of establishing clearly distinguishable male and female groups based on these measures.

The initial measure under investigation was head circumference, a widely used and easily obtainable anthropometric variable. Numerous studies have documented this measure, making it readily accessible for research purposes. In a study examining head circumference among adults in the UK, it was estimated that the average head circumference for males was 57.6 cm while for females, this was and 55.1 cm (Bushby et al., 1992). To determine an appropriate sample size, a power analysis was conducted, which indicated that 16 participants would be sufficient to establish distinct male and female groups.

Custom-fitted IMGs (IMGs) had been purchased for previous studies featuring players from both the university's men's and women's first rugby teams (E. M. P. Williams et al., 2021). These were used to measure the mean lateral (Y) component of translation for both male and female wearers. These involved measurements using calibrated digital callipers to determine the distance between canines and 1st molars. Statistically significant disparities were discerned between the male and female groups and are reported in Table 8-1. Additionally, a positive correlation was observed between the inter-tooth measurements and head circumference. Based on a power analysis, sample sizes of around 22-26 participants were deemed suitable for these measurements. However, when considering anthropometric measurements from a previous study, (such as height, mass, and head circumference), the predictions consistently indicated a participant count of fewer than 20 (E. M. P. Williams et al., 2021).

The same mouthguards were used to calculate angles of rotation, measured by the angles formed by the outer edges of the canine and lateral incisor, and the outer edges of the 1st molar and 2nd premolars. These measurements were conducted using an online tool and photographs taken from directly above the IMG at a consistent position and angle. These showed no significant differences between male and female groups

and were not used to determine sample size. These angle and distances are shown within Figure 8-1.

Figure 8-1: An image of an instrumented mouthguard, with the angle at the canine (orange) and 1st molar (yellow) highlighted, along with inter-canine distance (top, red) and inter-molar distance (bottom, red).

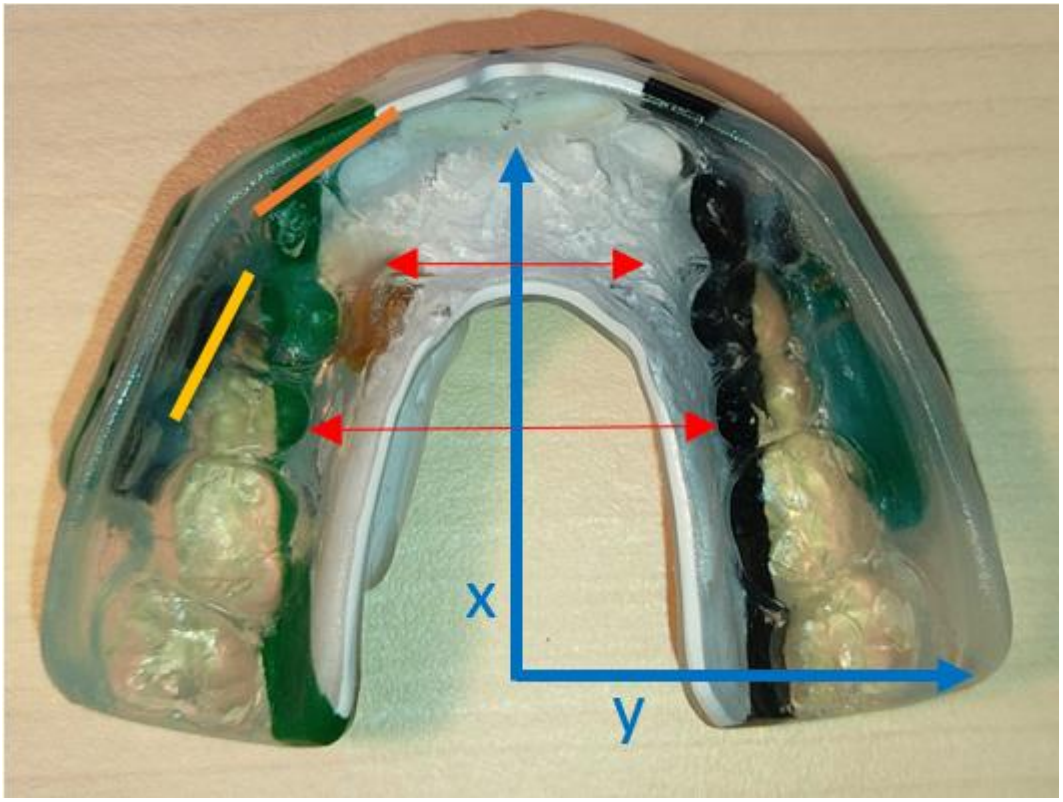


Table 8-1: Dimensions of dentist-fit Prevent Biometrics IMGs.

Sex	Measurement Type	Distance at 1st Molar (mm)	Angle of 1st Molar (degrees)	Distance at canine (mm)	Angle of incisors (degrees)
Male (n=15)	Mean	55.3	22.4	38.2	59.1
Male (n=15)	Standard deviation	2.8	3.9	2.3	5.0
Female (n=15)	Mean	52.2	22.9	35.4	59.2
Female (n=15)	Standard deviation	2.0	2.8	1.5	4.5
	Difference in means	3.1	0.5	2.8	0.1
	p-Value	0.0013	0.749	0.00055	1.0

8.2.2 Participants

Following the power analysis results, a cohort of 25 participants was recruited specifically for this study, comprising 11 females and 14 males, none of whom participated in previous studies within this thesis. The inclusion criteria stipulated that participants had to be at least 18 years old and engaged in regular physical activity. Written informed consent was obtained from each participant before their involvement in this study. The participants underwent a series of procedures, including anthropometric measurements, dental impressions, and 3D optical scans of the head. Prior to the start of this study, ethics approval was obtained from Swansea University ethics committee, ensuring compliance with institutional guidelines (ethical approval number FP_01-09-22).

8.2.3 Anthropometric Data

Height, Mass & Head Circumference

Anthropometric data were collected using methods compliant with the International Society for the Advancement of Kinanthropometry (ISAK) guidelines (Silva & Vieira, 2020). The measurements taken from the participants were mass, height, and head circumference of the participant. All measurement equipment was calibrated prior to use if required. Other measurements were taken from the participants, being the lip thickness, a dental impression, and a 3D optical scan of the participants' head. The collection of these measures is detailed below.

Lip Thickness

The thickness of the upper lip was required to be measured so that the sensor locations could be estimated from the optical scans. A wooden tongue depressor was given to each participant to create an estimate of the thickness of the lip. The tongue depressor was gently pressed against the front teeth and held in place by the participant. With the tongue depressor in this position, the participant was asked to mark with a pencil on the tongue depressor as closely to the lip as possible without compressing the lip. This would allow an estimate of the lip thickness by measuring the distance between the end of the tongue depressor and the pencil mark.

Dental Impressions

Participants were asked to create an impression of the upper dentition, which could later be used to aid the estimated location of the kinematic sensors within the wearers head. Each participant was provided with a dental impression tray containing a premixed dental alginate. They were then asked to place this securely on their upper dentition for the required time as stated in the alginate instructions, before being removed and the impression being left to set. Each impression was then disinfected and left to set, before being stored in an airtight container, labelled with an ID number given to anonymise each participant. Photographs were later taken of all impressions, each photo was taken from the same distance and with the same camera for each mouthguard, with the impression adjacent to a known scale for measurements to be taken accurately.

The distance for the Y component of translation was measurable from the dental impression. As sensors in commercial devices are located by the incisor or molar, the Y component of translation for molar placed sensors could be estimated by measuring the distance from the estimated sensor location to the midpoint of the impression. The intra molar distances were measured from the outer edge of the tooth. The Y distance for the translation could then be reported as this measurement divided by two. This provided the distance between the sensor and the centre of the head, on the assumption that the teeth are symmetrically aligned in the head, with the centre of mass acting as the central (medial) plane. The precision of this measurement will depend on the thickness of the mouthguard. Because the sensor is positioned on the outer of the device, the measurement may be affected by a few millimetres, depending on the thickness of the mouthguard material.

The angles that the teeth formed relative to the medial plane were also measured for the purpose of aligning the directions of X and Y with those presented in Figure 8-1. This involved establishing a reference point at 0 degrees along the medial plane (reference the same image as previous sentence). Subsequently, an angle was determined by creating a plane that intersected the edges of the teeth surrounding the point of interest and measuring the deviation from this reference point. These angle measurements were specifically taken at the 1st molar and the incisor to ensure accuracy and alignment with the designated dental locations.

8.2.4 Optical Scan of Head

A 3D scan of each participant's head was acquired using a Creaform HandySCAN 3D optical image scanner (Creaform, Québec, Canada), offering mesh resolution of down to ± 0.1 mm. To streamline file management and improve scanning efficiency, the device was calibrated to operate at an accuracy setting of 0.5 mm. additionally, further calibration was implemented to prior to scanning each participant to optimise the scanners performance for their skin tone.

An issue present with scanning head is the presence of the participant's hair. As the hair masks the shape of the head and isn't relevant to the head in terms of this project it is important to reduce it effect on the calculation of the CG. To do this, participants were asked to wear a tight-fitting elasticated cap, to which targets were adhered to aid the scanner in mapping the head, shown in Figure 8-2. This improved the accuracy of the optical scanner and pull the hair close to the scalp to give a better representation of skull size. For participants with long hair a cap with a whole at the crown was used so that hair could be styled outside of the cap so that it would be easily removable from the optical scan STL file in post processing. An unprocessed scan is show in Figure 8-3, prior to removing these features.

Figure 8-2: A participant undergoing head scanning, with the yellow cap to compress the hair visible. Black and silver markers are visible on the cap to aid the scanner to map the geometry of the head.

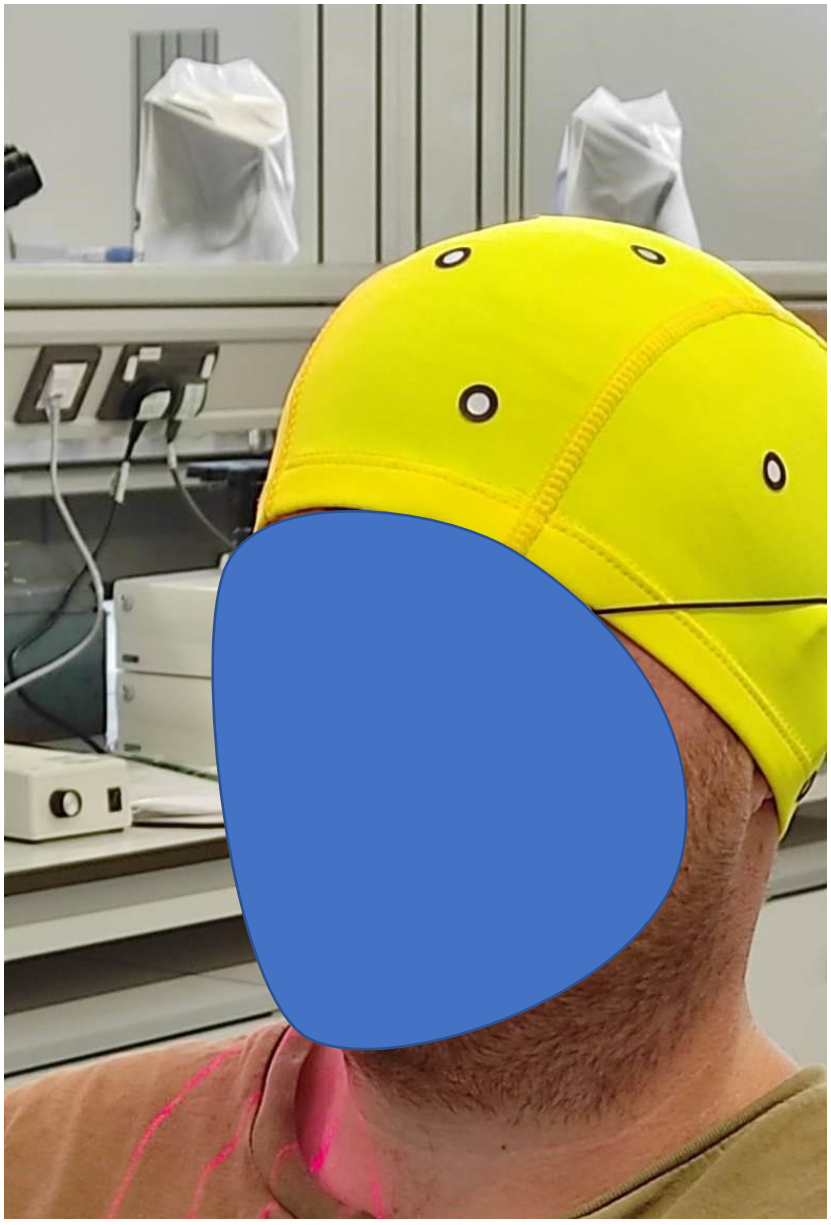
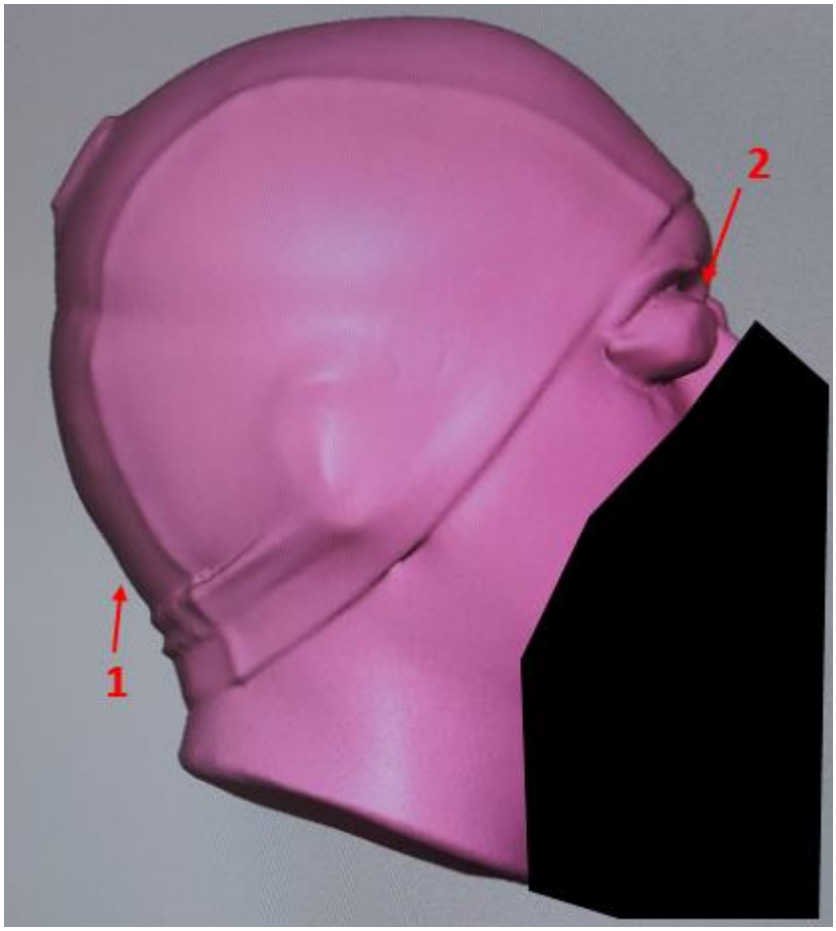


Figure 8-3: A scan of a head with cap (1) and eye protection goggles (2) labelled.



To enhance precision during the scanning process, the optical scanner was passed around the participant's head to produce a 3D optical scan. Subsequently, the obtained 3D model was imported into a solid modelling computer-aided design software called Freeform (Artec 3D, Luxembourg). Within this software, post-processing tasks were performed, including the removal of features such as goggles used for eye protection, neck elements, and excess hair.

Additional markers were strategically placed on the participant's face, neck, and shoulders to enhance precision during the scanning process. These markers played a critical role in assisting the optical scanner in accurately mapping the surface and correctly placing detected surfaces in their intended positions. While surfaces with many distinct features in close proximity may require fewer markers, they are essential in featureless areas, allowing the scanner in piecing together the surface with precision.

The scans were then cropped in accordance with the methodology employed in the study by (Yoganandan et al., 2009), removing areas of the digital object that were below and behind the mandible. Subsequently, the centre of mass was located by finding the digital objects geometric central point as determined from the now processed optical scan. Using the estimated centre of mass of the head, distances were computed from the gap between the lips along the estimated medial plane, to the geometric centre of mass of the head. This provided values of X, Y, and Z. in relation to these measurements, locations of sensors could be estimated by combining the lip thickness measures and the dental impression.

8.2.5 Measuring Sex Differences

To estimate the effect of these measures, t-tests were conducted in Python for each of the anthropometric measures, between the male and female groups. The purpose was to determine whether there were distinctions between the male and female groups for the standard anthropometric measures. This would then enable a quantitative evaluation of how these sex differences related to the estimated measurements mapping the sensor to the head CG location estimated with Freeform. Additional correlation coefficients and R^2 values were calculated between each of the measures taken. This was done to assess the relationships between the measures to see whether simple measurements i.e. height could be used to more accurately predict translation distances in order to make translations simpler.

Head acceleration data from genuine HAEs were then plotted, using the measurements from each of the participants to translate the linear acceleration. The resultant acceleration was calculated at the estimated CG for each participant using the rigid body translation. Box plots were used to visually represent the impact of translation on maximum resultant linear acceleration compared to the linear acceleration magnitude recorded by the sensor. These plots displayed the interquartile range (IQR) within coloured boxes, with the whiskers extending to 1.5 times the IQR. This was plotted for 16 example impacts which were taken from data collected from a women's rugby match. For these16 impacts, the resultant acceleration at the head's CG was estimated using the X, Y, and Z values estimated from the head scans. These impacts were selected at random, and the magnitude was estimated at the CG of each participant.

8.3 Results

A full set of anthropometric measurements from each of the 25 participants were collated, and the statistical summary for the male and female groups given in Figure 8-1. Among these nine measurements, significant differences were observed between the male and female groups in five key metrics: height, body mass, neck circumference, head circumference, and head volume. Due to unexpected shrinking of the dental alginate measurements could not be taken from the dental impressions and therefor an assesment of the effect of placing a sensor on the molar could not be calculated.

To explore the relationships between these anthropometric measures, a correlation plot and R-squared plot were created (Figure 8-4, Figure 8-5). The correlation coefficients ranged from 0.77 to -0.68, with the strongest correlations observed between mass and neck circumference (0.77), neck circumference and sex (0.73), and head circumference and volume (0.72). The correlations between the X, Y, and Z measurements were also examined, finding that height displayed the highest correlation with these variables, with correlation values of 0.25 and 0.23 with X and Z respectively. The R-squared plot was used to illustrate the strength of the trends between features. The relationship with the highest R-squared were the neck circumference and mass (0.59) followed by neck circumference and sex (0.53), volume vs mass/sex both scored the same value (0.47).

In the analysis of the 16 impacts, shown in Figure 8-6 - Figure 8-9, it was observed that the resultant acceleration for males and females typically exhibited similar group-level values. However, substantial differences often existed between the minimum and maximum values recorded within each group. For instance, in impact M, the maximum male impact reached 26.21 *g*, while the minimum was 21.39 g. For the female group, the maximum impact was 25.61 *g*, with a minimum value of 19.40 *g*. This revealed a variation of 27.59% within the female group, with a 20.25% difference between the maximum and minimum male values and a 29.86% disparity between the maximum male and minimum female impacts. Not all impacts displayed such significant variations, however. For example, impact L showed only a 5.68% variation within the female group and a 2.45% variation within the male group. These within-group variations were notably smaller compared to the variation observed in the sensor-recorded linear acceleration, which exhibited a percentage difference exceeding 85%.

Notably, the sensor magnitude often appeared as a significant outlier, distanced from the transformed data points.

Table 8-2: Statistical summary of male and female anthropometric measures.

Measurement	Female Mean	Female Standard Deviation	Male Mean	Male Standard Deviation	T-Statistic	P-Value
Height (cm)	168.03	6.49	177.89	5.03	-3.03	0.00594
Mass (kg)	67.28	7.68	79.20	6.49	-3.82	0.00089
Neck circumference (mm)	324.73	1.52	366.50	1.87	-5.06	0.00004
Head circumference (mm)	550.36	1.03	579.21	1.57	-3.73	0.00109
Head volume (mm^3)	3718315.71	244,608.83	4,170,588.53	214,088.45	-4.53	0.00015
X (mm)	101.01	8.17	104.88	5.10	-1.26	0.22142
Y (mm)**	3.27	5.66	0.36	4.03	1.45	0.16176
Z (mm)	71.08	48.21	72.23	39.84	-0.32	0.75379

Figure 8-4: Correlation plot between anthropometric measures.

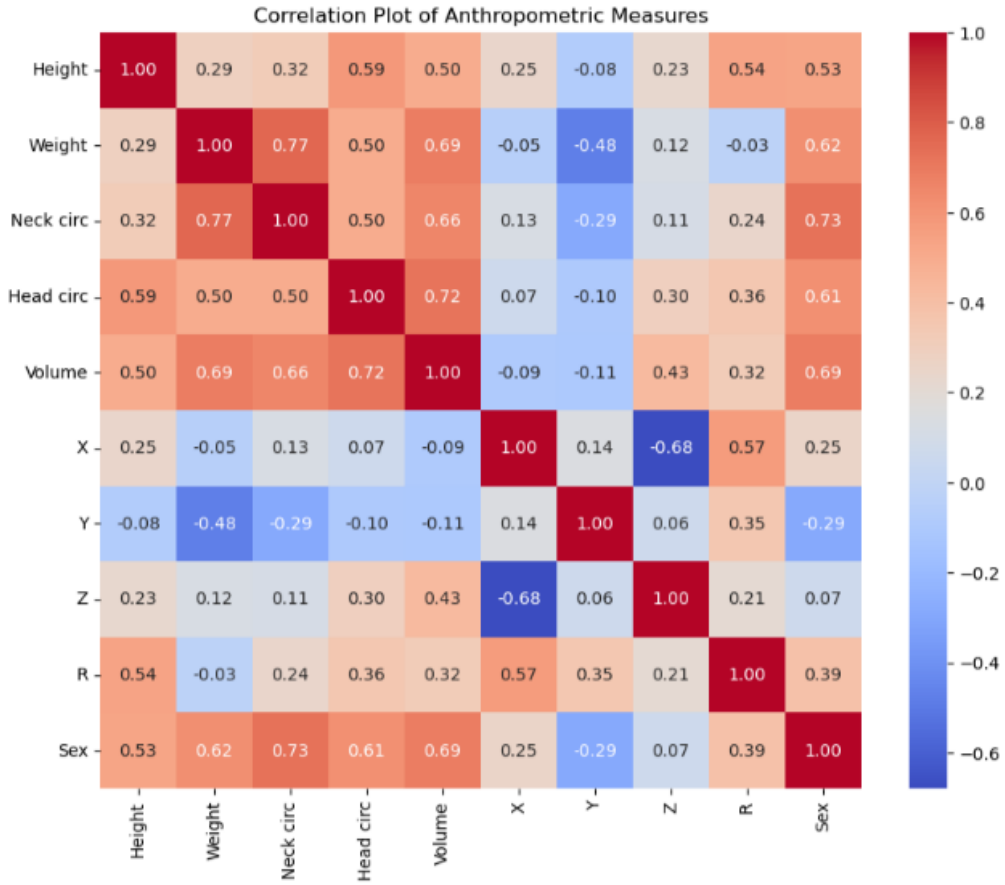


Figure 8-5: R-squared plot between anthropometric measures.

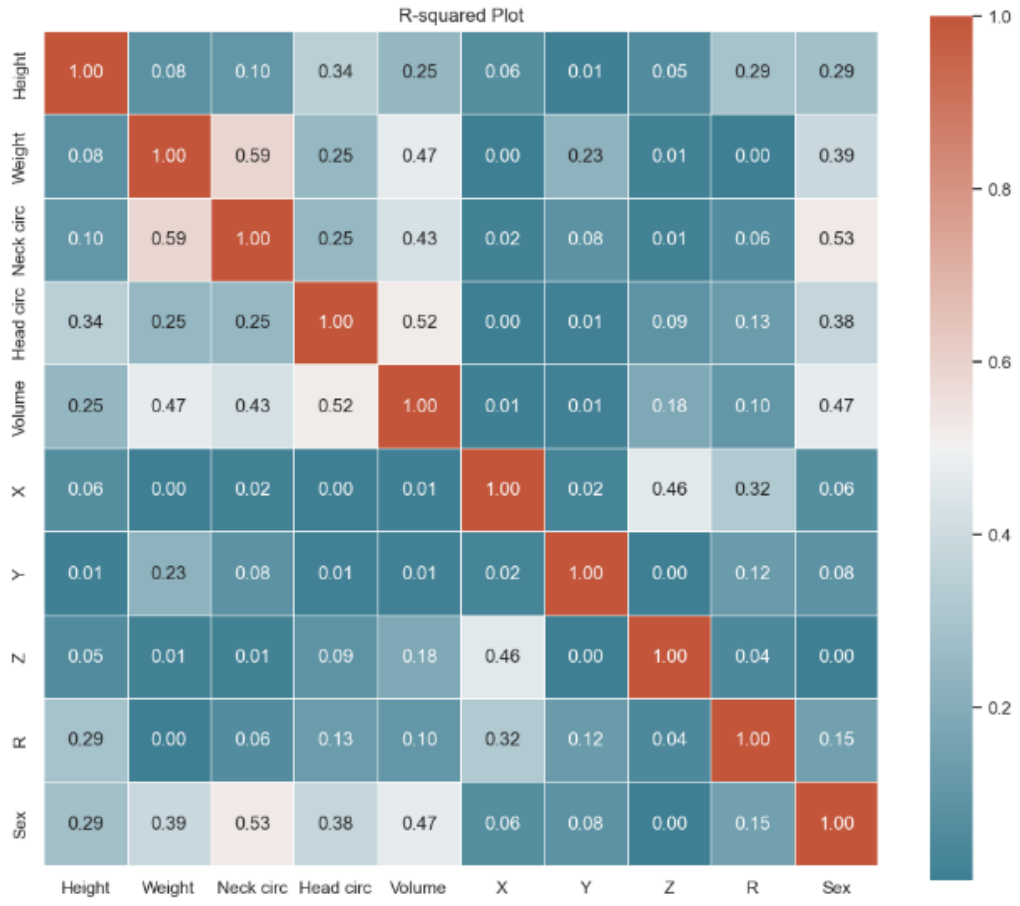


Figure 8-6: The peak linear acceleration of each impact after translation to the head's CG. 'a0' represents the linear acceleration at the sensor, with each point being the value after translation to a participants CG. Impacts labelled A-D.

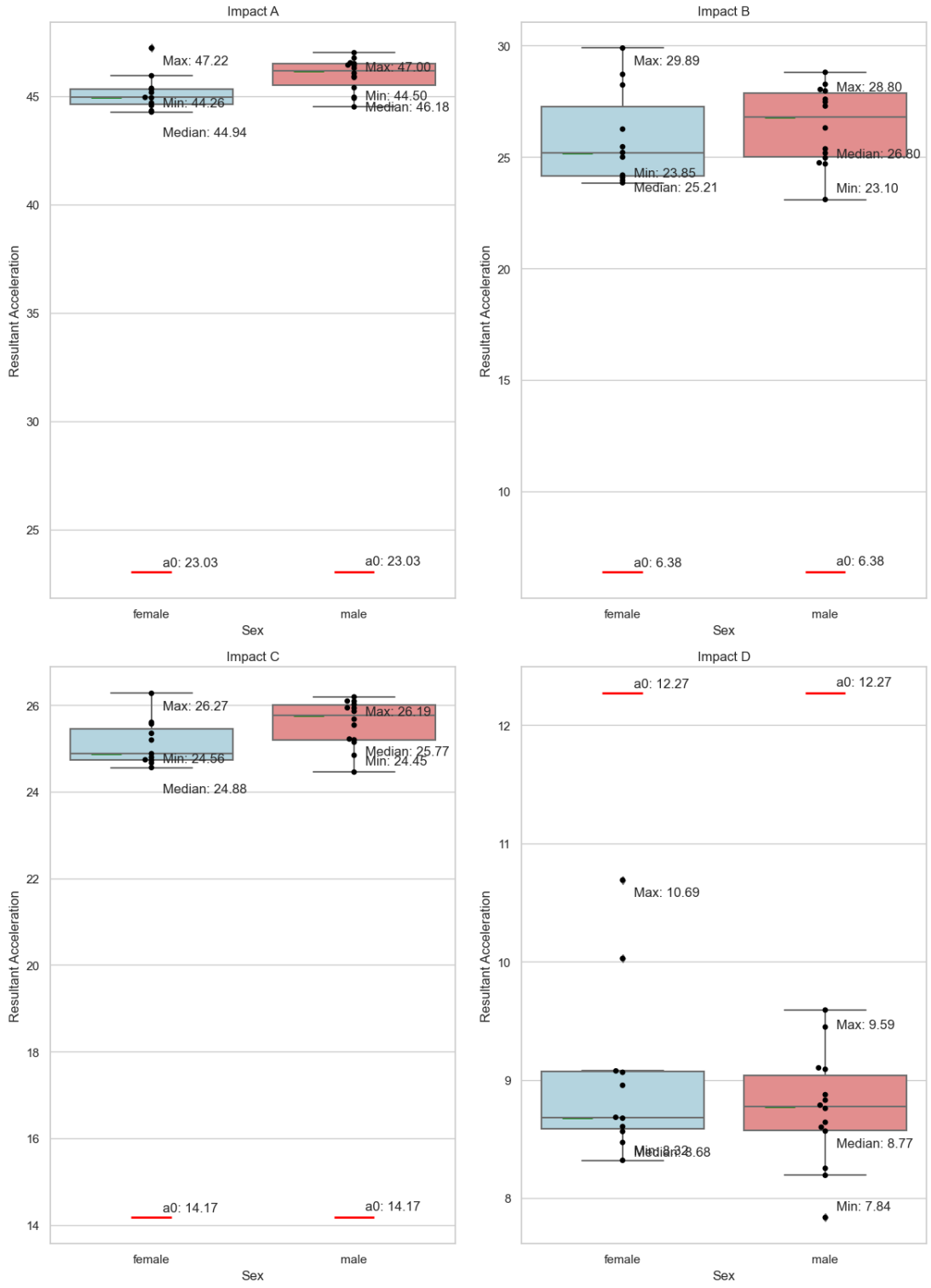


Figure 8-7: Impacts labelled E-H.

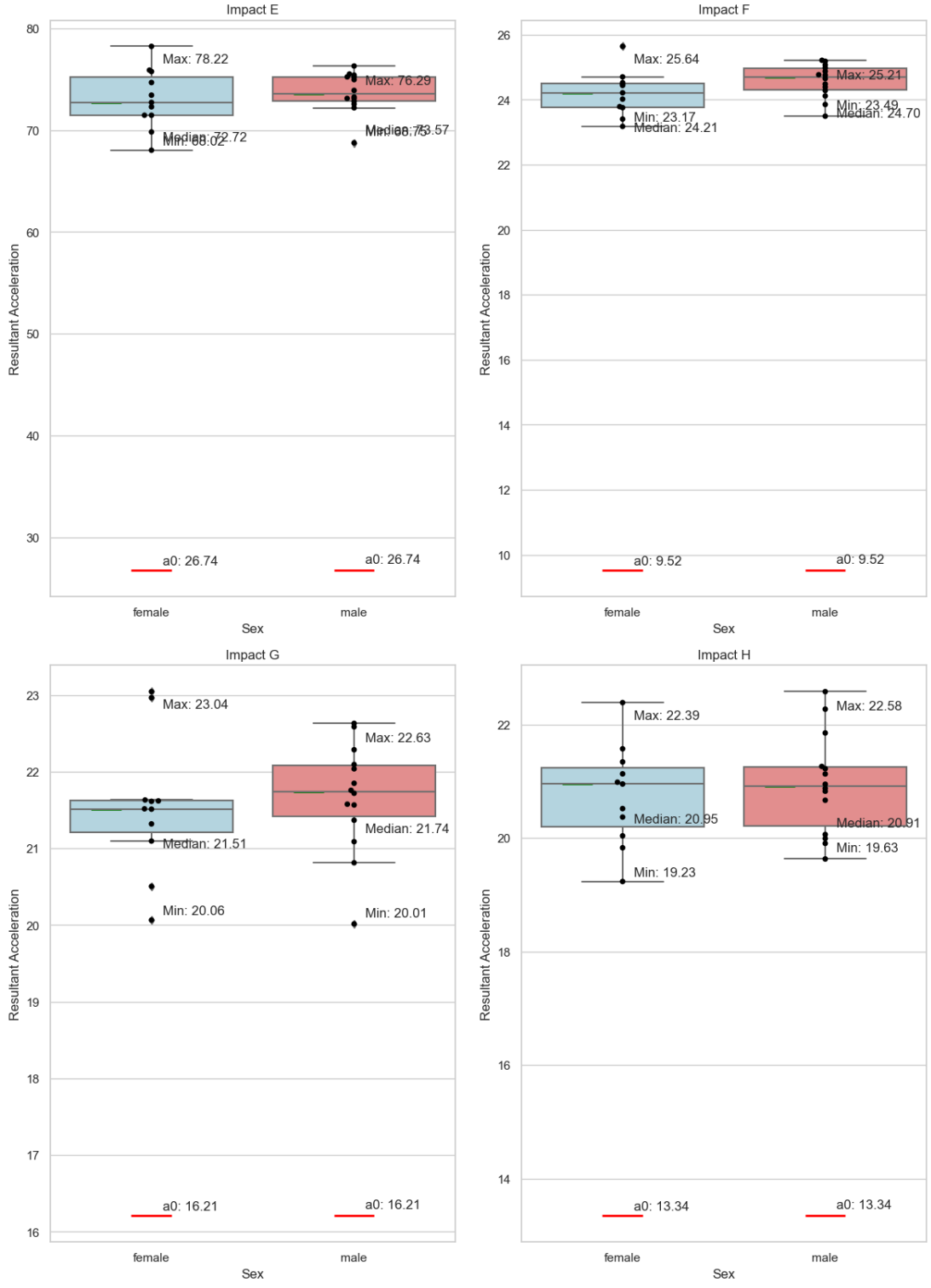


Figure 8-8: Impacts labelled I-L.

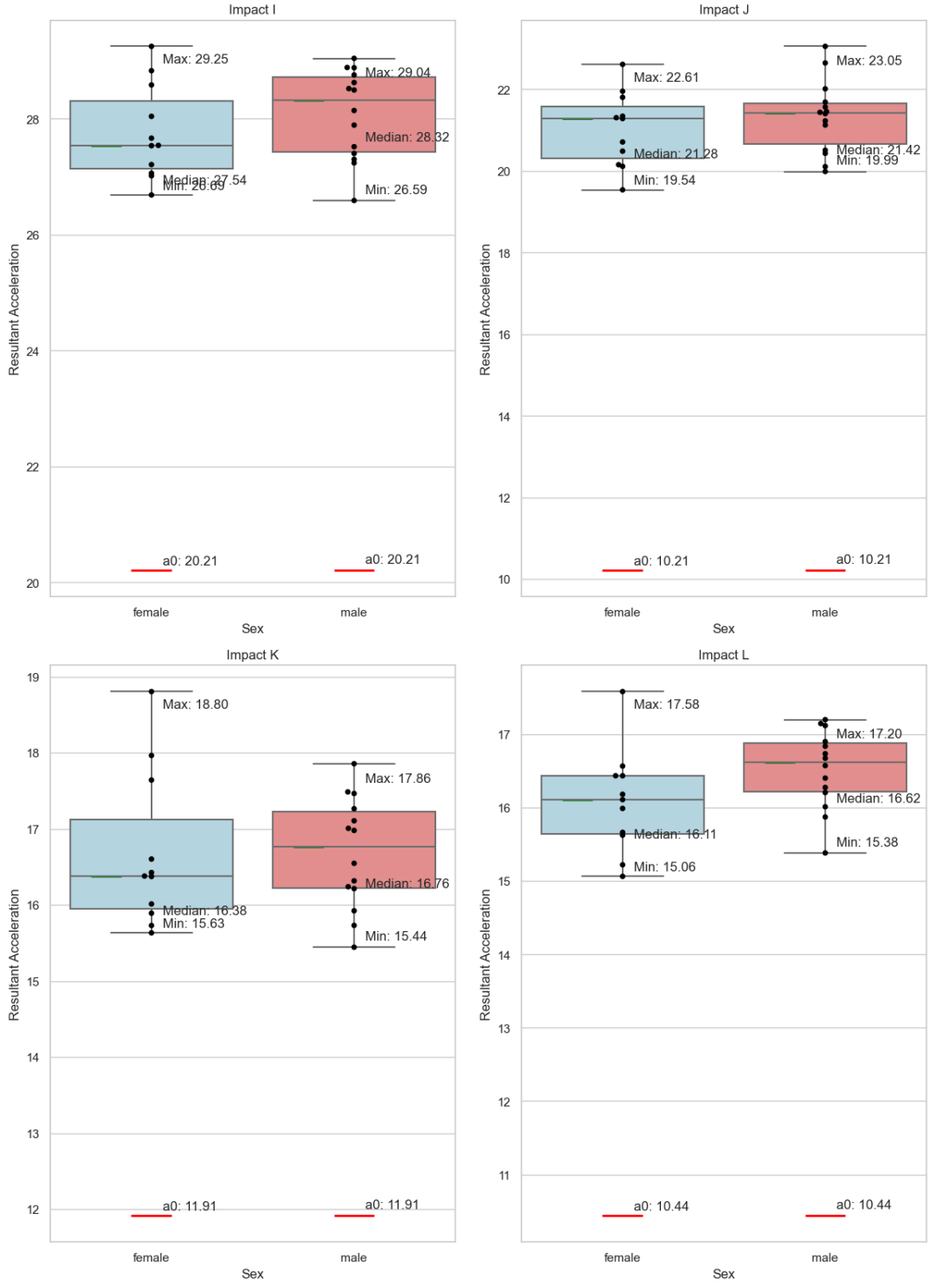
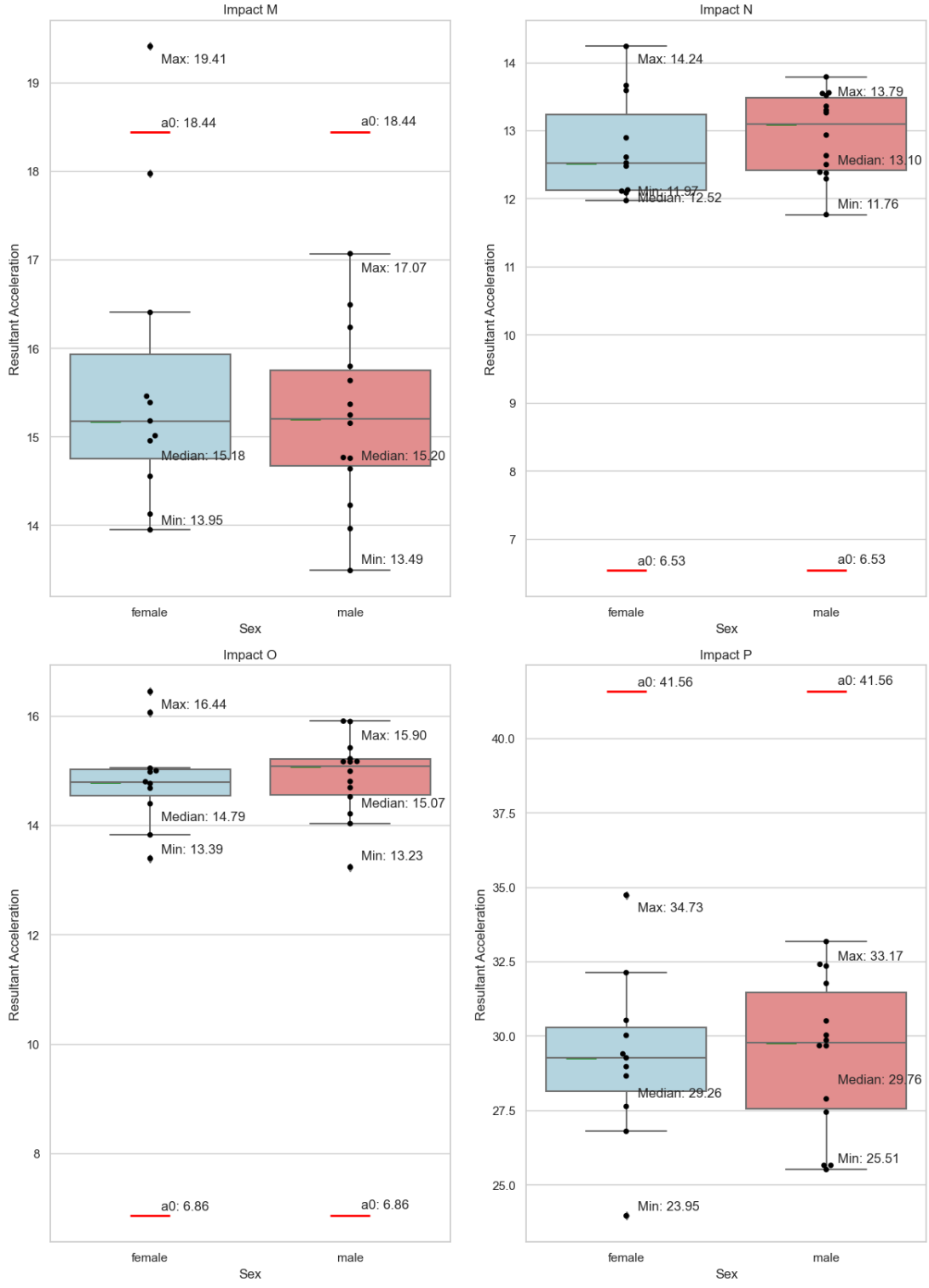


Figure 8-9: Impacts labelled M-P



8.4 Discussion

This study sought to examine the significance of the reporting location of head acceleration values with data recorded from IMGs, particularly the difference between

male and female heads. Anthropometric measurements were taken from 25 participants (14 female, 11 male) which were used to translate the linear acceleration from an estimated sensor location to an estimated head CG position. Kinematic measurements were taken from genuine HAEs gathered from women's RU matches. The process of collecting this data is descried within chapter 5. Plots were created to compare the maximum value of linear acceleration for 16 example HAEs. This is the first study to compare the effect of reporting the linear acceleration using personalised measurements and a male and female cohort. Previous studies have used CG measurements that are based off estimated 50^{th} percentile CG positions, derived from predominantly male cohorts (Bartsch et al., 2014; Yoganandan et al., 2009). This is the first study to assess the impact of translating linear acceleration to the head's center of gravity across various head sizes within both male and female cohorts.

This study achieved its goal of creating a comparison of how kinematics recorded at the sensor may vary between individuals when reported at the head's CG. The preliminary work for this study focused on how this may vary between sex, which led to the creation of a cohort that featured a distinctly male and female group, by a number of metrics. There were significant differences between, heights, mass, neck circumference, head circumference, and head volume between the male and female groups. Despite this there was no strong relationships between the translation distances between the sensor and CG and any other factor in this study. This means that sex or any of the anthropometrics were not seen to be an important factor when finding the distances between the sensor and the head CG. It was also seen that on most occasions, applying any translation to the data would provide more accurate results than reporting at the sensor. After applying a translation to the linear acceleration data, the next best way to improve the recording accuracy would be to apply a translation to the data that is more specific to the participant. Whilst occasions saw variations of over 85% between the sensor recording and median values, there were still variations of around 30% within the translated data.

This 30% variation in measurements illustrates that one size will not fit all in terms of applying a translation to linear acceleration data. Larger cohorts would likely include a greater range of head sizes shapes, which may result in even more pronounced differences between participants. Ensuring that these metrics could be improved and provide more accurate and precise data for each participant will help the field to

develop greater understandings of head accelerations in sport, and the effect on participants. With more reliably reported data this may then highlight particular trends in head acceleration in different participants that were not present if standardised translation values were used, e.g. translations to the estimated 50th percentile head CG. This may be particularly pronounced if devices were to be used within youth contact sports, as children with smaller heads will have inaccurate figures for head linear acceleration reported.

The study initially set out to determine measurements based on the estimated sensor location next to a molar and incisor. However, unexpected shrinkage of the alginate used for dental impressions rendered this exact goal unattainable. Consequently, the study shifted its focus to assess the impact of translating acceleration from the incisor only. This change had its advantages and challenges. This simplified the translation process by allowing the assumption of a y translation distance of 0 mm. This was useful as some asymmetries in the models and the cuts meant that attaining a ground truth model of the cranium only was difficult. Nevertheless, these small deviations, typically within a single millimetre, are unlikely to be of significant consequence. Additionally, measurements in the preliminary studies showed only discrepancies in the intra-molar distance of only a few millimetres, meaning that this would be unlikely to lead to large deviations between participants.

Future work could include using alternative methods such as medical imaging methods like CT or MRI scans to estimate the distances for translation for IMG data. As many of these images are taken daily there could be potentially a large cache of medical images that would be usable for the purpose of this study. It would allow for more accurate predictions of the distances involved as the sensor locations as they could be more accurately predicted within the head, rather than estimating using other measurements. Another addition that could make this line of study more valuable within the field would be the addition of more anthropometric measures with which to better predict the head CG in relation to the sensors position. For example, it possible to measure the outside of the head by hand to see whether the distances between the sensor location and head's CG could be reliably predicted with a simple measure (Yoganandan et al., 2009). Should this again not be viable, AI could potentially be used to better predict the translation distances for the participants using a number of anthropometric measures.

Regarding the field and translating data to the head's CG, a standardised and transparent procedure would play a pivotal role in facilitating the development of repeatable data transformations and analysis. This, in turn, would pave the way for more collaborative efforts in the analysis of HIT data. By establishing a common framework for data interpretation, sports organisations, researchers, and equipment manufacturers can work together to enhance player safety. Furthermore, such standardisation would provide a foundation for the development of new tools and technologies, ultimately aiming to improve their functionality and further elevate the safety measures in sports.

In conclusion, this study aimed to investigate the differences between device reporting location and the effect on reported values, and the effect of sex and anthropometrics may play on measurement accuracy. The study highlighted that reporting from the sensor will not provide consistent results in regard to acceleration at the head CG and that there can be large variations in the CG linear acceleration for real world impacts. Future work should focus on creating easy to measure metrics that can allow from transformation to be conducted quickly and reliable data to be reported allowing for comparison between devices from a large cohort.

With data validated and magnitudes corrected, the next stage of the pipeline outlined in Chapter 4 is to extract information relating to the causation of the HAE. This will be studied within the following chapter.

9 Automated Epidemiology of HIT Data

9.1 Introduction

Brain injuries cause acute and chronic neurocognitive damage in athletes (Nowinski et al., 2022). Sporting governing bodies are becoming increasingly concerned about the incidence of severe brain injury and are implementing interventions to improve athlete safety (Pfaller et al., 2019; Raftery et al., 2021; Rutherford et al., 2019). The rate of diagnosed brain injuries has been used as a metric to measure the effectiveness of interventions in previous studies (Cosgrave & Williams, 2019; Mack et al., 2021; Stokes et al., 2021). Because of the complexity of brain injuries, this may not be the optimal approach, as symptoms can present subtly and require a subjective assessment (McCrory et al., 2017).

Instrumented mouthguards can provide a quantitative source of information to support clinical decision-making and diagnostics (Kieffer et al., 2020; Patton, Huber, Jain, et al., 2020). Although IMGs cannot diagnose brain injury, these data can measure the effectiveness of interventions (Arbogast et al., 2022). It was recommended that video footage is used to verify the data recorded by the IMG is accurate verification (Patton, Huber, Jain, et al., 2020). Although video review is a time-consuming process, it can provide more reliable data and offers the opportunity to analyse the causation of the head acceleration events (HAEs) verification (Patton, Huber, Jain, et al., 2020). Researchers have developed ML algorithms capable of automatically validating HAEs, with these algorithms achieving comparable performance to video review verification (Gabler et al., 2020; Goodin et al., 2021; Miller et al., 2018; Raymond et al., 2022; Wu et al., 2018).

Despite the success of these algorithms, no publications detail using ML to predict the causes of HAEs. Researchers have extensively studied the use of ML algorithms to identify human movements from sensor data, a field known as action recognition. (Dang et al., 2020). Optical, kinematic, and environmental sensors have been used to detect and analyse human movements in various settings, including assisted living, security, and sports analysis (Dang et al., 2020). In sports analysis, studies have used kinematic sensors to identify movements in sports, such as volleyball, ballet, and rugby (Gastin et al., 2014; Hendry et al., 2020; Kautz et al., 2017).

This study investigates whether action recognition can detect the causes of head acceleration events in women's RU using the kinematic data recorded by IMGs. The study focused on two tasks: identifying whether HAEs arose from direct contact with the head (DHC) or without direct head contact (NDHC) and distinguishing between events recorded by the BCs and tacklers. Successful implementation of these methods will greatly reduce the time burden when investigating the causes of HAEs in RU and could offer real-time insights into athlete health and performance.

9.2 Methods

9.2.1 Data Collection

This study uses data collected and processed in Chapters 5 and 6, where full descriptions of the processes involved are reported

9.2.2 Classifier Selection and Training

The study undertook two classification tasks. The first task was distinguishing between HAEs resulting from DHC and NDC and the second task distinguished between HAEs recorded by tacklers or BCs. A dataset was created for each task, with Figure 9-1 showing a comparison of the aggregated linear acceleration for the data of both classes in each task. This figure illustrates that there are notable differences between the acceleration profiles of the data, which may help to develop appropriate features to capture these differences.

This data was then split into training and testing groups for both classification tasks. The training groups contained approximately 80% of the data, with the remainder used as a test set. Data were split so that the entirety of recordings from each session would either appear in the training or test set. Each dataset underwent the same feature transformation, creating features predominantly derived from studies creating HAE classification algorithms (Gabler et al., 2020; Goodin et al., 2021; Wu et al., 2018). The feature categories included wavelet transformations, pulse parameters, PSD measures, and statistical measures. More details of these features can be found within Chapters 5 and 6, with deeper descriptions of the underlying methods within Chapter 3.

Figure 9-1: A comparison of the linear and rotational acceleration profiles for DHC head impacts and NDHC head accelerations.

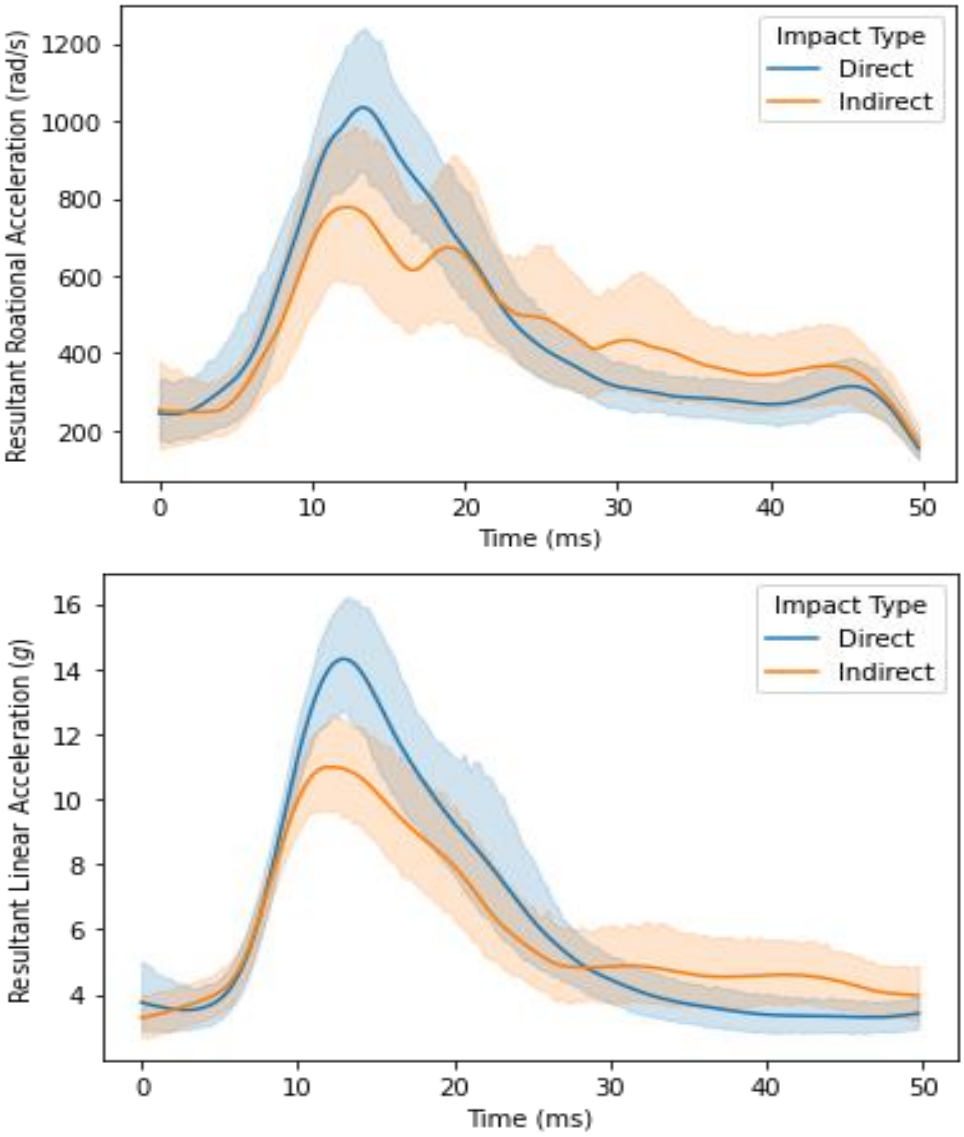
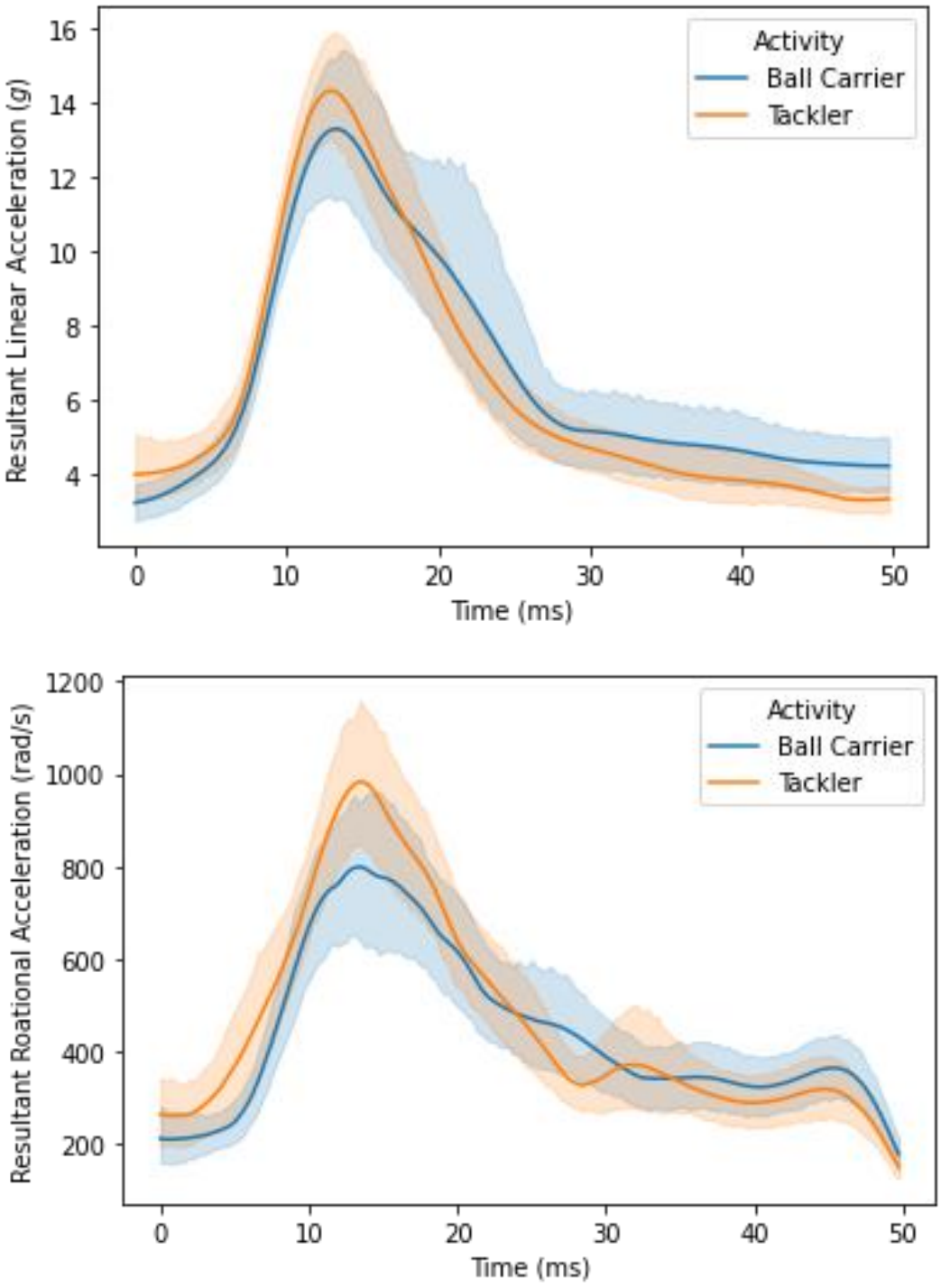


Figure 9-2: A comparison of the linear and rotational acceleration profiles for head accelerations recorded by BCs and tacklers.



Data Pipeline

The statistically derived FCQ variant of the mRMR was used to reduce the feature count (Zhao et al., 2019a). Further details on this method can be found in 3.3.2, while the methods used to reduce the features are reported in Chapter 6.

With the initial feature count reduced, an isolation forest was used to identify and remove outlying data points from within the training dataset, to improve the generalisability of the training dataset (Liu et al., 2008b). Isolation forests are covered in more detail in Section 3.3.3. After outlying data were removed, the data were then scaled such that for each feature the data had a mean of zero and a standard deviation of one.

The next stage was to further reduce the feature count by 50%. For classifiers capable of calculating feature importance, recursive feature elimination was used to reduce the feature count, allowing for the selection of features that directly impacted model performance. In recursive feature elimination, a model is selected and trained with the entire dataset, the feature importances are calculated, the feature with the lowest importance is removed. This process was repeated until the desired number of features had been achieved. For classifiers that were incapable of calculating feature importance, feature counts were reduced using the mRMR method again.

Two classifiers used in this process were logistic regression and an MLP (3.4). Feature importance was not considered in the MLP algorithm. Before the data pipelines, hyperparameter optimisation was undertaken for both classifiers, using ten-fold cross-validation with ten descriptive features. Logistic regression was optimised through the testing of multiple C values, penalties, and solvers whilst the number of layers and neurons per layer were tested for the MLP. For each architecture tested for the MLP, the layers would contain the same number of neurons.

Once hyperparameters were selected, the training pipeline was initiated first with two features, increasing in increments of two, up to 50. This meant that both classifiers were trained 25 times, with the number of features ranging from one to 25, after feature selection. Due to the small dataset and the amount of collinearity between features, 25 total features was selected as a maximum value, as it was close to the lower limit suggested by Hua et al., 2005(Hua et al., 2005)**.** This allowed for classifier-specific feature optimisation later in the data pipeline. Once the algorithm was trained, the features and their coefficients were used to transform the test dataset. The trained classifiers then predicted data classes for the test dataset.

The metrics used to assess the classifiers' performance on the training and test data were the macro recall score and the area under the receiver operator curve (AUROC).

Recall (Equation 3-50) is defined as the percentage of labels that were assigned to the correct class, whereas macro recall is the unweighted average of the recall scores for each class (8-2). Further details on AUROC can be found in section 3.5.2.

$$Macro\ Recall = \frac{Recall_0 + Recall_1}{2} \quad (9\text{-}1)$$

Because of the stochastic nature of the training process and the small dataset size, the pipeline was repeated 100 times per feature number, with each iteration initiated with a random seed value from 0 to 99 to create a repeatable pipeline. This allowed for measurement of both over-training and provided unbiased estimates of the classifier's performance. Once the performance of the 100 iterations had been measured, the mean of the scores over 100 iterations was recorded, as were the highest scores. When reporting the highest scores, only classifiers which had a training and test difference of <2.5% were reported to avoid the inclusion of classifiers with large over-training. This figure was selected as an arbitrary measure of overfitting, as no specific measures are available within the field to define an overfitted classifier.

9.3 Results.

Action Recognition Study

From the video review, 169 events (65 BC, 104 tackler) were identified for this study, which were split into training and test sets. The test set contained 34 events (12 BC, 22 tackler) with 135 within the training set (53 BC, 82 Tackler events). Data was split so that no session data appeared within both the training and test datasets, although data from the same participants may appear in both. For binary classification, negative class labels were assigned to BC events, with positive labels given to tackler events. Features were reduced from an original count of 603 to 50, with the mRMR method. This included the features derived from each feature category. A total of 18 wavelet transformations, 14 power spectral densities, and nine pulse parameters were used with the remaining features coming from derivatives, PCA stats and area methods. Features crafted from each inertial measure and direction were used in the feature set.

On the training data, the logistic regression classifier achieved a macro recall of 69.3% and AUROC means of 0.753 when trained at 24 features. When classifying the test data, the highest scores recorded by the logistic regression classifier occurred with 20

features, with macro recall and AUROC means of 69.4%, and 0.721, respectively. Regarding individual classifier performance, the highest was achieved by the logistic regression classifier of 73.5% for macro recall and 0.765 for AUROC respectively, with the training and test scores within 1% of each other. The results are summarised in Table 9-1.

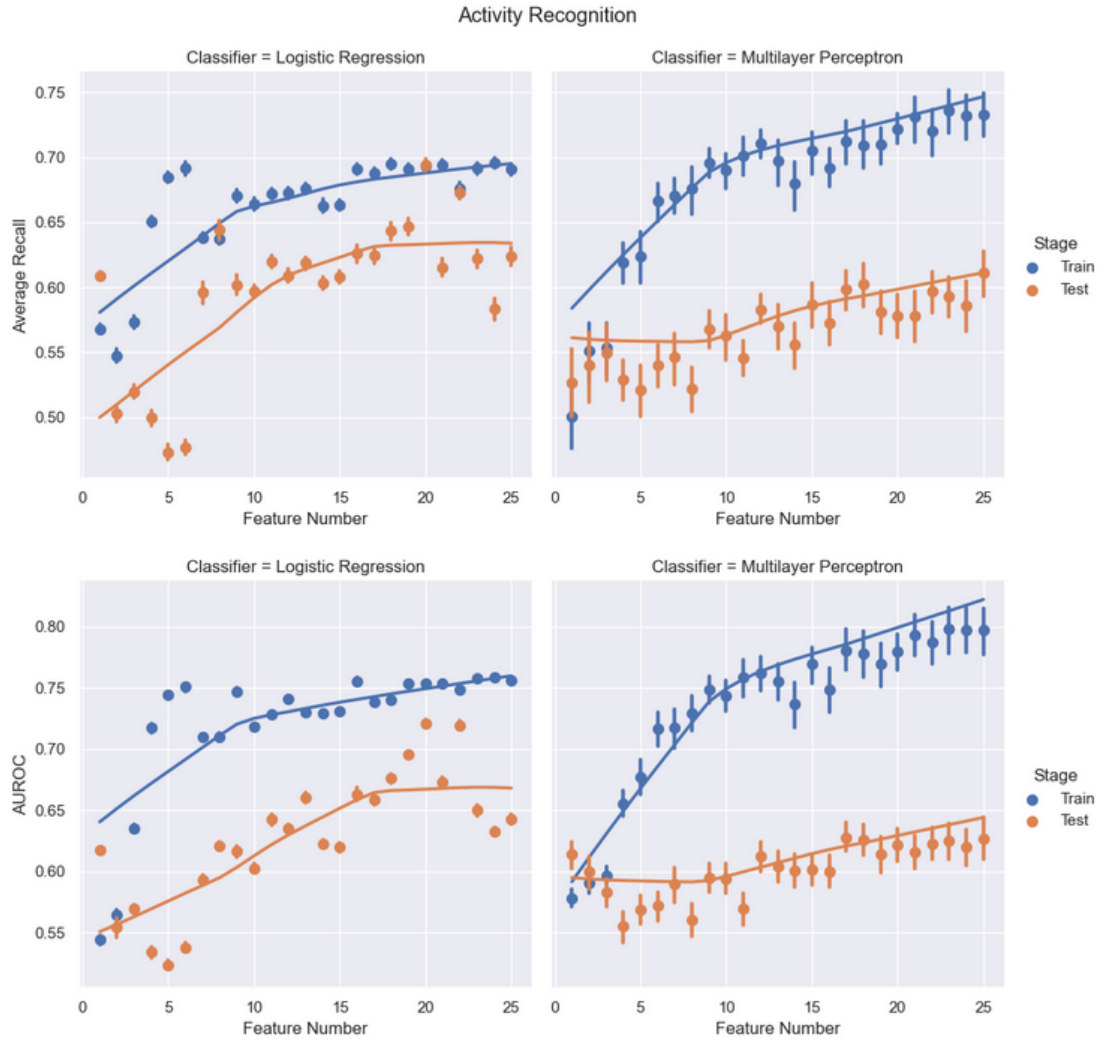
The MLP classifier achieved macro recall and AUROC means of 73.6% and 0.798 on the training dataset. On the test data, the MLP classifier achieved a peak means of macro recall and AUROC of 61.1% and 0.628. The highest performance of individual classifiers was recorded at 87.9% and 0.795 for macro recall and AUROC respectively on the test data. All results may be seen in Table 9-1.

Table 9-1: Action Recognition Results

Action Recognition		Average – Macro Recall	Average – AUC	Max* – Macro Recall	Max* – AUC
Logistic Regression	Training	69.3%	0.753	72.2% 20/54	0.754 20/25
	Test	69.4% (20F)	0.721(20F)	73.5%	0.765
Multilayer Perceptron	Training	73.3%	0.797	80.7% 24/69	0.784 12/59
	Test	61.1% (F25)	0.628 (17F)	81.6%	0.795

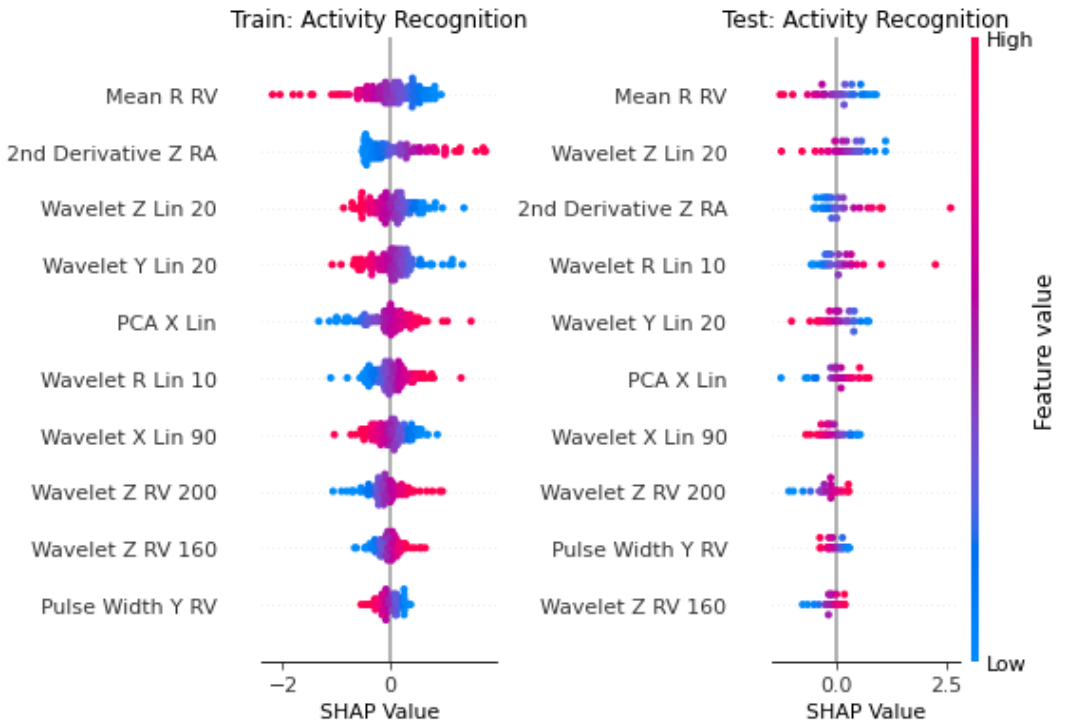
The logistic regression classifier saw increases in training scores for AUROC and macro recall as the number of features increased. However, there was a marked reduction in the rate of improvement after 15 features. This trend was mimicked by the test data, although the scores were typically lower than those of the training data. The MLP classifier tended towards 75% and 0.8 for macro recall and AUROC as the number of features increased in the training phase. The testing scores of the MLP saw less improvement, achieving scores in the low 0.6s for both. The standard deviation of the classifiers over the 100 repetitions was much lower for the logistic regression than the MLP, as illustrated in Figure 9-3.

Figure 9-3: Results for the action recognition classification task. Showing the classifier performance vs. feature count for the logistic regression classifier (left) and multilayer perceptron (right), with the training (orange) and test scores (blue) shown.



SHAP analysis was conducted on the highest-performing iteration of the logistic regression. This was conducted on both the training and test sets to inspect the consistency of feature importance between the two groups. Of the top ten features, the most important features were wavelet transformations recorded from all inertial measures and all orientations. Frequencies used ranged from a minimum of ten Hz to a maximum of 200 Hz, with frequencies closer to these limits more prevalent. The most important feature for training and testing was the mean resultant rotational velocity, with high values predictive of BCs. The summary of the SHAP values can be seen in Figure 9-4.

Figure 9-4: SHAP values: Action Recognition. Feature names express the feature group, recording axis (x/y/z/resultant), kinematic measure (linear acceleration, rotation velocity/acceleration) and recording frequency in Hz.



Impact Type

An impact-type training dataset was created with 91 impacts (68 DHC and 36 NDHC), with a test set of 33 recordings (23 DHC, ten NDHC). The feature count was reduced to 50 features with mRMR to remove low-performing features and improve training times. The 50 features selected contained 25 wavelet transformation features, with the remaining 25 made of all other feature categories. Negative class labels were assigned to DHC impacts and positive class labels were given to NDHC impacts. Data was split so that data from the same session would not appear in both the training and test groups, although data from the same athlete may.

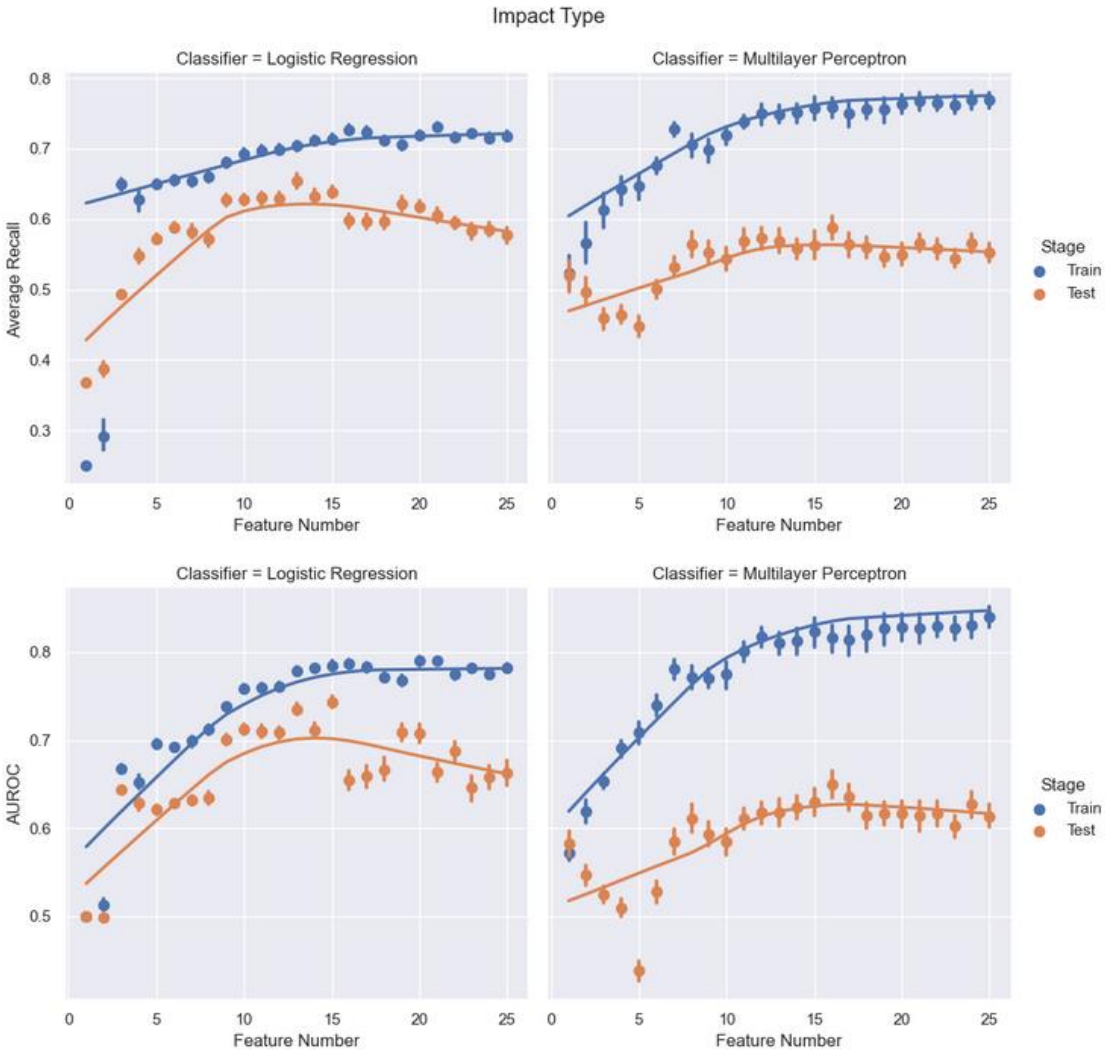
On the training data, the logistic regression classifier achieved macro recall and AUROC means of 70.4% and an AUROC of 0.785. When classifying the test data, the highest scores recorded by the logistic regression classifier were macro recall and AUROC means of 65.4%, and 0.744, respectively. In terms of the performance of individual classifiers, the highest performance achieved by the logistic regression classifier on the test data was 74.6% for macro recall and 0.813 for AUROC. All results for both classifiers are shown in Table 9-2.

Table 9-2: Impact Type Results

Impact Type		Average – Macro Recall	Average – AUC	Max* – Macro Recall	Max* – AUC
Logistic Regression	Training	70.4%	0.785	73.9%	0.805
	Test	65.4% F13	0.744 F15	74.6% 13/84	0.813 15/62
Multilayer Perceptron	Training	76.0%	0.816	76.5%	0.803
	Test	58.8% 16F	0.649 16F	77.5% 24/62	0.791 12/58

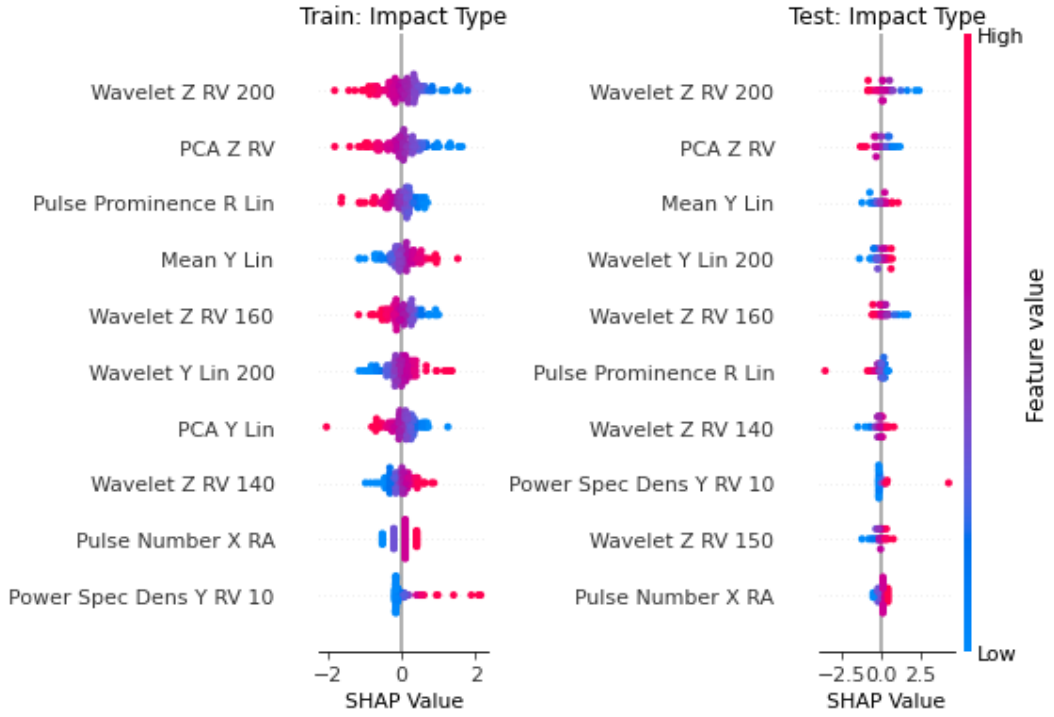
The logistic regression classifier saw increases in the training scores for AUROC and macro recall as the number of features increased to 15, before plateauing for both metrics. A similar trend was seen in the test results, with some decline after ten features were used to train the logistic regression classifier. This is indicative of overfitting for these high feature count classifiers. The MLP classifier saw similar trends in performance as the feature count increased, with the rate of improvement decreasing between ten and 15 features for predictions of the training labels. The multi-layer perceptron classifiers showed erratic performance between one and ten features, but marginally improved after 11 features. The standard deviation of the MLP classifier over the 100 repetitions was smaller than that of the previous classification task. However, it was still greater than the logistic regression classifier. The results are displayed in Figure 9-5.

Figure 9-5: Results for the impact type classification task. Showing the classifier performance vs. feature count for the logistic regression classifier (left) and multilayer perceptron (right), with the training (orange) and test scores (blue) shown.



SHAP analysis was conducted on the logistic regression classifier that provided the highest individual score for AUROC while within 2.5% of the training score. Features used included a range of all feature classes, with features relating to high frequency more common. The most important feature for both the training and test was a wavelet transformation of linear acceleration at 200 Hz. The PCA of the Z direction rotational velocity and the mean Y linear acceleration were identified as similarly important. For the impact detection task, the SHAP results can be seen in Figure 9-6.

Figure 9-6: SHAP values: Impact Type. Feature names express the feature group, recording axis (x/y/z/resultant), kinematic measure (linear acceleration, rotation velocity/acceleration) and recording frequency in Hz.



9.4 Discussion

In this study, two binary classification tasks were attempted. The first task was an action recognition algorithm, predicting whether the impact was measured in the BC or tackler. The second task was the creation of an impact-type detection algorithm, predicting labels for DHC and NDC. This study was conducted to investigate whether more aspects of the video verification process could be automated. While multiple studies have reported success in automating impact verification, no studies have reported automated action recognition or impact causation from IMG data. In this study, mean test scores of macro-recall and AUROC of 65.4%/0.744 and maximum scores of 77.5%/0.791 were achieved for the impact type detection classification task (Table 9-2). For the action recognition classification task, mean test scores of macro recall and AUROC of 69.4%/0.721 were achieved, with individual classifiers recording highs of 81.6%/0.795 (Table 9-1).

No studies have reported using HIT devices for action recognition, making direct performance comparisons impossible. Some of the closest studies have classified motions in ballet and RU, reporting higher classification performance (Chambers et al., 2019b; Hendry et al., 2020). Chambers et al., (2019) classified one-on-one tackle

and ruck events using IMUs attached to the upper back, reporting an accuracy of 100% (Chambers et al., 2019b). Hendry et al., (2020) classified ballet motions, with classification accuracy ranging between 75.1% and 97.8% using an array of six sensors (Hendry et al., 2020). Whilst the performance of the classifiers in this study was lower, it shows that classifiers can be created with IMG data for impact causation analysis.

Action Recognition

From the action recognition study, wavelet transformations were identified by both SHAP and mRMR as the most predictive feature category. This included features derived from each type of motion and each orientation. A diverse range of other features was included, both those typically associated with impact verification and the features derived from specific action recognition studies (Chambers et al., 2019b; Gabler et al., 2020; Goodin et al., 2021; Hendry et al., 2020; Wu et al., 2018). From the calculated SHAP values, it was shown that frequencies closer to the feature extremities appeared more often than those in the middle of the range (Figure 9-4). The differences between the predictions of the features were also subtle. For example, 20 Hz linear acceleration was more predictive of the tackler class, whilst ten Hz linear acceleration was predictive of the BC class (Figure 9-4).

Mean resultant rotational velocity was the most predictive feature in both training and testing, with high mean rotational velocity values more predictive of BC events (Figure 9-4). This is perhaps due to the mechanics of tackling in RU. BCs often transition from an upright running position to a prone position on the ground during a contact event. As the BC falls, there is a period where the player accelerates towards the ground under gravity, allowing for the head to achieve a higher rotational velocity. In comparison, the tackler's head is likely to be proximal to the BC, with the torso lowered. In this position, there is less potential to achieve high rotational velocity before a physical limit to movement is encountered and velocity is impeded. Greater linear acceleration in the 20 Hz band was predictive of BC events, slightly above the limits of voluntary human motion (Khusainov et al., 2013). As the primary contact point on the BC is often further from the head compared to the tackler, this suggests that the body may attenuate impact forces before significant head acceleration occurs, resulting in lower-than-expected head acceleration values. Conversely, high-frequency components were generally more indicative of events involving the tackler.

This may be attributed to the closer proximity of the tackler's head to the point of contact, which allows for less high-frequency attenuation and, consequently, greater transmission of these frequencies to the sensor.

Impact Type

For the impact type, the most common feature group selected by the SHAP and mRMR methods were wavelet transformations. Wavelet transformations represented 25 of the top 50 features. Of the wavelet transformations, higher frequency transformations from rotational velocities were associated with the NDHC class (Figure 9-6). Feature performance was less repeatable between training and testing compared to the action recognition study, although the top 10 contained the same features (Figure 9-6). This indicates that the trends found by the classifiers were not as consistent between training and testing when compared to the action recognition study. The strongest feature was the 200 Hz wavelet transformation in the Z direction for rotational velocity transformation, with high values predictive of DHC events (Figure 9-6). Vibrations of this frequency are transmissible through bone, which may indicate vibrations are transmitted directly to the accelerometer in DHC impacts (Chang et al., 2018). In contrast, similar frequency wavelet transformations of linear acceleration were more predictive of NDHC events. Given that inertial measurement units (IMUs) can capture noise associated with rapid exhalation, such as when wearers shout, it is plausible that similar noise occurs when air is expelled from the lungs during contact events. This phenomenon may be more common in non-direct head contacts, where head acceleration originates from torso impact, potentially inducing forced exhalation.

Classifier Performance

These studies achieved average recalls of >64% in both classification tasks, which is lower than other action recognition tasks (Chambers et al., 2019b; Hendry et al., 2020). Several factors are likely to have contributed to this. Only a small dataset was available for each task, reducing the capacity to identify the important patterns. From an action recognition perspective, performance improvements may occur by increasing the length of the recording windows of each event. While a range of recording frequencies has created effective action recognition classifiers (Awais et al., 2016; Csizmadia et al., 2022), recording windows >one second could reduce classifier performance in action recognition tasks (Banos et al., 2014; Csizmadia et al., 2022). Given that the duration of the recordings was 50 ms^{-1} and significantly shorter than those in

comparable studies, it is likely to be detrimental to classifier performance (Chambers et al., 2019b; Dang et al., 2020; Hendry et al., 2020). One of the most significant confounding factors was the inherent nature of the classification tasks examined. In each case, two athletes were involved in an impact, resulting in similar kinematic responses. Consequently, the ability to distinguish between measurements from the BC versus the tackler, or to determine DHC versus NDHC, was limited by the similarity in kinematic responses to these events.

Of the two classifiers, the logistic regression classifier achieved the highest classifier macro recall and AUROC means (Table 9-1, Table 9-2). The MLP scored the highest on individual scores (Table 9-1, Table 9-2). As the MLPs can form large networks, the optimisation algorithm may create many local minima. Because of the small datasets and relatively large network used in these studies, the seed values became important when trying to converge on the global minima. Smaller networks could be used, however, these did not prove as effective when optimising the hidden layer size and number in the hyperparameter optimisation stages. The effect of the large network is illustrated by the over-training shown in Figure 9-3 and Figure 9-5 as the MLP performs better on the training data compared to the test data. The effect of the random seed across the can be seen by the size of the error bars in Figure 9-3. As a result, the random seed affected the results of the MLP more than those of the logistic regression algorithm, which converged on more similar values for training and test set classification scores.

Feature Optimisation

SHAP and mRMR were used to reduce the feature count, with both methods largely agreeing on feature importance. As the feature count was increased from one to 15, training scores improved in both tasks. Beyond this point, overfitting was evident as performance measures declined (Figure 9-3, Figure 9-5), suggesting that additional features contributed less to classification accuracy. The steady performance increases up to the 15-feature mark indicates that repeatable patterns are present within the dataset.

Conclusions

Previous studies have created HAE classifiers and action recognition models. In these tasks, ML algorithms have been successful in achieving human-level performance (Gabler et al., 2020; Goodin et al., 2021; Raymond et al., 2022; Wu et al., 2018). No

existing studies have attempted to combine approaches and automate the action recognition from HIT data. The current study has illustrated that it is possible to create classifiers to detect actions or impact types using IMGs. Future research in this field will allow for the development of more powerful classification algorithms to improve the performance in these classification tasks. The development of HIT devices with recording windows more similar to those used in action recognition studies will also likely improve classifiers' performances within these tasks. Such improvements may facilitate more effective and informative indicators of athlete performance and well-being.

This concludes both the pipeline outlined in Chapter 5 and the methods, results, and analysis aspects of this thesis. In the preceding chapter, the process and data collection and transformations have been reported. This processed data has then been used in aid of three novel studies, which offer methods to improve both the reliability of IMGs and functionality. The next chapter concludes this thesis, describing the conclusion, limitations, and future work.

10 Conclusions, Limitations, and Future Work

In this thesis, computational methodologies were used to enhance the reliability and functionality of IMGs, addressing prevalent issues in the field of HIT. The primary research question was whether the detection of authentic HAEs in female RU could be automated. To achieve this, a cutting-edge ML algorithm was developed to identify head accelerations events within female RU, yielding AUROC and AUPRC scores of 0.92 and 0.85, respectively. This algorithm presents a twofold advantage by enabling instantaneous detection of genuine events, as opposed to relying on video review and facilitating reliable identification of genuine impacts by researchers lacking prior experience. Consequently, future work can be conducted rapidly, reliably, and by less experienced operators, thereby mitigating existing obstacles that impede progress within the sports brain injury field.

The second research question investigated whether ML could automatically generate reports of HAE. To address this, two ML classification algorithms were developed to predict distinct characteristics of HAEs. These purpose-built algorithms were designed to differentiate between BCs and tacklers, as well as discern direct from indirect head contacts. Despite being developed using a limited dataset, both classifiers demonstrated promising outcomes, and have the potential to revolutionise the information extraction processes associated with HIT. Monitoring head accelerations to ascertain the precise mechanisms can be time-consuming and resource intensive. By further refining these classifiers, significant reductions in these requirements can be achieved, providing actionable insights into the causes of HAEs.

The third research question investigated the appropriate methods for reporting linear acceleration values. The objective was to investigate the effect of reporting the linear acceleration from differing locations, to determine whether the currently employed methods are adequate. The findings revealed substantial variations in the results between participants, suggesting that the reported data may lack consistency and accuracy between studies and individuals. By implementing the approach outlined within this study, the data can be reported more accurately for each individual, promoting consistency across studies.

The methodologies developed in this research represent significant advancements in HIT studies within the context of RU. These studies introduce innovative approaches

to enhance athlete safety through the automated detection and categorisation of HAEs, whilst improving the accuracy of the reported values. By providing these approaches, the processing time and effort required for data analysis can be reduced, allowing more individuals to actively engage in this form of research. This is of particular importance in light of the World Rugby IMG mandate implemented on the 1st of January 2024. The challenge known as the 'trillion sensor problem' pertains to the efficient utilisation and meaningful interpretation of the vast amount of data generated by IMGs and other sensors. Without a well-defined plan and adequate tools for data processing, the potential value derived from these data will be limited. The longer the lack of action persists, the more significant the issue becomes, and valuable insights are forfeited. The methods presented in this thesis have been specifically designed to address these risks, offering solutions that contribute to safer sports for all.

10.1 Specific Challenges

Initially, the intention of this project was to use a brand of IMG, referred to as IMG-A. As the thesis research questions began to take shape, considerable effort was made to develop solutions for the specific challenges encountered when using IMG-A. The considerable hardware performance issues encountered with this brand, however, were followed by unexpected limitations in the availability of these devices. These issues impeded any further progress and restricted the scope of any further investigation using the IMG-A brand, so the research transitioned to the Prevent Biometrics hybrid IMG.

An additional limitation of this PhD study was the impact of the COVID-19 pandemic, which rendered it impossible to collect data for an extended period within the timeframe of this thesis. This circumstance was particularly problematic given the suboptimal performance of the IMG-A devices and the lack of high-quality data they provided. As a consequence, the earliest available data for the IMG recordings used in this thesis was recorded on the 24th of November, 2021, highlighting a period of over two years before suitable data for the final projects could be obtained. Had the early data collection have succeeded and the pandemic been avoided it would have greatly increased the data available to us and perhaps strengthened some of the conclusions, particularly in the causation analysis studies. It would also have allowed for a greater comparison between males and females and between devices, which would have

opened up new research avenues. Unfortunately, there is not much that can be learned from the pandemic, other than to always have a backup plan in place, regardless of how safe first choice options appear and to make the most of the opportunities it presents. With no requirements for data collection it afforded time to read background literature, improve coding and analytical skill, and develop project plans, all of which meant that research and analysis could be effectively conducted once the regulations had lifted.

During this period, efforts were made to establish collaborations to obtain additional data for this project. Here are two notable examples: Following discussions with Prevent Biometrics, approximately 8,000 recordings containing spurious events from American football HAE recordings were acquired. The objective of acquiring these data was to enhance the algorithms developed within this project and develop an additional classifier specifically tailored for multi-sport HAE classifications. However, due to limitations in the data processing methods and the absence of accompanying video footage or event descriptions, the data proved unsuitable for integration into the algorithms. Furthermore, access to raw data itself was not available, further impeding its usability for this project.

Similarly, collaboration was sought with World Rugby's ORCHID project. Whilst a placement was agreed during the project, this was abandoned due to COVID-19 and associated New Zealand visa and entry restrictions. Thus, the data from the ORCHID study did not become available within the timeline of this PhD. Should this have transpired, video footage and data for >10,000 recordings would have become available for this thesis. This extra data would have greatly benefitted the work done for this thesis and expanded the applicability of the findings.

10.2 Future Directions

With the recent World Rugby mandate, IMGs have been touted as the device to revolutionise contact sports and bring a new layer of protection from brain injury. Whilst this is a lofty ambition, the exact mechanisms that will lead to this change are not clear. Devices can record accurate data with the correct algorithmic support, but these data are not enough to ensure player safety alone. Much more research needs to be conducted to offer explanations to these data, and data needs to be interpreted to create an understanding of how and why it occurred. As the interpretation of the data

is the most valuable stage, and also the most time consuming, efforts should be made from within the industry to reduce this effort and add value to the data, to allow for actionable insights to be derived off the back of the analysis. The question becomes how high quality data can be collected and reported quickly, easily, and reliably. A number of steps have been outlined in the following discuss that will offer direction in achieving the maximum value from IMGs. This work initially builds upon the work outlined within this thesis and offers suggestions for future projects.

In this thesis we developed an algorithm to predict genuine and spurious HAEs, which achieved cutting-edge performance in classifying events in female RU. Algorithms such as this will continue to achieve higher performance and become more generalisable to new data should the training dataset continue to grow. Therefore, in the pursuit of higher performing algorithms, datasets should continue to be added to with new, verified data in order to achieve this better performance. The more reliable these algorithms become, the higher the quality of data will be collected, and the more valuable insights will be found. Whilst adding more data will create better classifiers, should the data be of a certain quality, the field could see improvement by looking to create more generalisable algorithms. If algorithms are tailored to specific systems, sexes, ages, and sports then there will be a vast number of algorithms to fit combinations of these factors that need to be created. Therefore, in the interest of simplicity, creating algorithms that use data from diverse datasets containing examples from a variety of these factors should be created and reported. This may have adverse performance results, but it may also lead to the creation of more powerful algorithms and should enough data be collected allow for the use of more powerful ML algorithms.

Action recognition has been addresses within this thesis with the creation of two algorithms that aim to predict information around how a HAE occurred. Without algorithmic support this stage will be fully reliant of analysts reviewing video footage and linking it with head accelerations to document these mechanisms. The time and expense associated with this analysis could ultimately result in no or incomplete analysis. This could lead to missing vital insights, which may be avoided by the successful implementation of these algorithms. To improve these algorithms there are two parts that need to be addressed. Firstly, the acquisition and use of more data within the training will allow for the detection of more concrete trends that allow for better

performance. Secondly, the identification of the most relevant facets of play to be detected will allow for more targeted algorithms to detect them. Addressing these points will allow for these algorithms to make the greatest impact in the field. This could additionally be supported using algorithms harnessing the data of other wearable sensors, commonly worn by athletes, which are well adapted to the detection of activity.

Within this thesis the variation between linear accelerations recorded at the sensor and at the estimated CG of various wearers was highlighted. As IMG systems do not factor this into the reported linear acceleration it means that the results are not a truly representative of the linear acceleration experienced by the wearer at the head's CG. The work within this theses allowed for more accurate prediction of the linear acceleration, however the methods were too time consuming to be appropriate to use with many wearers. To build on this work, more results need to be collected so that a distribution of head measurements are available, to inform researchers as to how these measurements will vary. Additionally, these measurements need to be tied to an easier to collect anthropometric measure, so that it is feasible for all researchers to be able to accurately report linear acceleration. As little correlation was seen between measure such as height and weight, more detailed physical measurements of the head may provide opportunities to quickly improve data quality. With rugby being a game for all, it is important that this study is not limited to specific cohorts, and should aim to gather data from diverse cohorts.

The ultimate aim of the IMGs would be the detection of concussive injury, or the detection of a near-dangerous level of sub-concussive injury. This is a complicated challenge due to the complexity and subtlety of brain injury, but steps may be taken towards it. Firstly, the development of targeted features will be required, as current studies have not achieved concussion prediction. These features will likely require more time-dependant characteristics, such as measures that describe the accumulation of impacts With initiatives such as World Rugby's recent mandate, it should lead to datasets containing vast numbers of HAEs, from both injury causing events, and non-injurious events. Classification algorithms may then provide an opportunity for the automation of injury prediction. By using similar methodologies to those used within this thesis a classifier may be constructed using the HAE data. For example, training a classifier with HAEs that led to injury and HAEs that did not, algorithms will be

capable of predicting a probability of a new event belonging to either group. This would allow for the first steps in creating a ‘smart’ mouthguard capable of detecting injury. To extract more information from these data, combining it with the explainable AI methods used within this thesis will allow for a greater understanding of what does and does not characterise a brain injury.

10.3 Conclusions

The work within this thesis lays a foundation for advancements in the domain of IMGs and HIT in sports, particularly in the context of enhancing safety measures in RU. The work conducted in this thesis offer solutions to make IMGs more valuable now and pathways to continue improving them for future research. This continual development will allow for a future where data-driven insights can revolutionise injury prevention strategies. However, this potential can only be fully realised through continued efforts. By augmenting IMG systems with algorithmic support, diversifying training datasets, and fostering collaborations across research spheres, these devices hold the promise of fundamentally reshaping the landscape of contact sports safety. The journey toward creating a safer environment for athletes is ongoing, and if the momentum of this thesis's work is sustained, these devices can undoubtedly play a pivotal role in significantly reducing the risks associated with head injuries, ultimately making rugby and other contact sports a far safer and more secure arena for athletes.

11 Bibliography

Aagaard, J. S., & Du Bois, J. L. (1962). Telemetering impact data from the football field. *Electronics*, *36*, 46–47.

Aitkin, M., & Foxall, R. (2003). Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, *13*(3), 227–239. https://doi.org/10.1023/A:1024218716736

Alpaydın, E. (2013). Introduction to Machine Learning. In *MIT Press* (Vol. 19, Issue 2). https://doi.org/10.1017/s1351324912000290

Alsalaheen, B., Landel, R., Hunter-Giordano, A., Shimamura, K. K., Quatman-Yates, C., Hanke, T., & McCulloch, K. L. (2019). A Treatment-Based Profiling Model for Physical Therapy Management of Patients Following a Concussive Event. *Journal of Orthopaedic & Sports Physical Therapy*, *49*(11), 829–841. https://doi.org/10.2519/jospt.2019.8869

An, T. K., & Kim, M. H. (2010). A new Diverse AdaBoost classifier. *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*, *1*, 359–363. https://doi.org/10.1109/AICI.2010.82

Arbogast, K. B., Caccese, J. B., Buckley, T. A., McIntosh, A. S., Henderson, K., Stemper, B. D., Solomon, G., Broglio, S. P., Funk, J. R., & Crandall, J. R. (2022). Consensus Head Acceleration Measurement Practices (CHAMP): Origins, Methods, Transparency and Disclosure. *Annals of Biomedical Engineering*. https://doi.org/10.1007/s10439-022-03025-9

Arshad, M. A., Shahriar, S., & Anjum, K. (2023). The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques - Case Of Breast Cancer Diagnosis. *ArXiv.Org*. https://doi.org/10.48550/ARXIV.2306.02449

Awais, M., Palmerini, L., Bourke, A., Ihlen, E., Helbostad, J., & Chiari, L. (2016). Performance Evaluation of State of the Art Systems for Physical Activity Classification of Older Subjects Using Inertial Sensors in a Real Life Scenario: A Benchmark Study. *Sensors*, *16*(12), 2105. https://doi.org/10.3390/s16122105

Azouvi, P., Arnould, A., Dromer, E., & Vallat-Azouvi, C. (2017). Neuropsychology of traumatic brain injury: An expert overview. *Revue Neurologique*, *173*(7–8), 461–472. https://doi.org/10.1016/j.neurol.2017.07.006

Bailes, J. E., Petraglia, A. L., Omalu, B. I., Nauman, E., & Talavage, T. (2013). Role of subconcussion in repetitive mild traumatic brain injury. *Journal of Neurosurgery*, *119*(5), 1235–1245. https://doi.org/10.3171/2013.7.JNS121822

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window Size Impact in Human Activity Recognition. *Sensors*, *14*(4), 6474–6499. https://doi.org/10.3390/s140406474

Bartsch, A. J., Samorezov, S., Benzel, E., Miele, V., & Brett, D. (2014). Validation of an 'intelligent Mouthguard' Single Event Head Impact Dosimeter. *SAE Technical Papers*, *2014-Novem*(November). https://doi.org/10.4271/2014-22-0001

Beckwith, J. G., Zhao, W., Ji, S., Ajamil, A. G., Bolander, R. P., Chu, J. J., McAllister, T. W., Crisco, J. J., Duma, S. M., Rowson, S., Broglio, S. P., Guskiewicz, K. M., Mihalik, J. P., Anderson, S., Schnebel, B., Gunnar Brolinson, P., Collins, M. W., & Greenwald, R. M. (2018). Estimated Brain Tissue Response Following Impacts Associated With and Without Diagnosed Concussion. *Annals of Biomedical Engineering*, *46*(6), 819–830. https://doi.org/10.1007/s10439-018-1999-5

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bergtold, J. S., Yeager, E. A., & Featherstone, A. M. (2018). Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *Journal of Applied Statistics*, *45*(3), 528–546. https://doi.org/10.1080/02664763.2017.1282441

Boswell, D. (2002). *Introduction to Support Vector Machines*.

Brady, T. (1991). Neural networks—an overview. *An Introduction to Neural Networks*, 9–12. https://doi.org/10.1201/9781315273570-6

Breedlove, E. L., Robinson, M., Talavage, T. M., Morigaki, K. E., Yoruk, U., O'Keefe, K., King, J., Leverenz, L. J., Gilger, J. W., & Nauman, E. A. (2012). Biomechanical correlates of symptomatic and asymptomatic neurophysiological impairment in high school football. *Journal of Biomechanics*, *45*(7), 1265–1272. https://doi.org/10.1016/j.jbiomech.2012.01.034

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broglio, S. P., Lapointe, A., O'Connor, K. L., & McCrea, M. (2017). Head Impact Density: A Model to Explain the Elusive Concussion Threshold. *Journal of Neurotrauma*, *34*(19), 2675–2683. https://doi.org/10.1089/neu.2016.4767

Buschjäger, S., Honysz, P.-J., & Morik, K. (2022). Randomized outlier detection with trees. *International Journal of Data Science and Analytics*, *13*(2), 91–104. https://doi.org/10.1007/s41060-020-00238-w

Bushby, K. M., Cole, T., Matthews, J. N., & Goodship, J. A. (1992). Centiles for adult head circumference. *Archives of Disease in Childhood*, *67*(10), 1286–1287. https://doi.org/10.1136/adc.67.10.1286

Bussey, M. D., Davidson, P., Salmon, D., Romanchuk, J., Tong, D., & Sole, G. (2022). Influence of the frame of reference on head acceleration events recorded by instrumented mouthguards in community rugby players. *BMJ Open Sport & Exercise Medicine*, *8*(4), e001365. https://doi.org/10.1136/bmjsem-2022-001365

Bussey, M. D., Salmon, D., Romanchuk, J., Nanai, B., Davidson, P., Tucker, R., & Falvey, E. (2023). Head Acceleration Events in Male Community Rugby Players: An Observational Cohort Study across Four Playing Grades, from Under-13 to Senior Men. *Sports Medicine*. https://doi.org/10.1007/s40279-023-01923-z

Caccese, J. B., Buckley, T. A., Tierney, R. T., Rose, W. C., Glutting, J. J., & Kaminski, T. W. (2018). Sex and age differences in head acceleration during purposeful soccer heading. *Research in Sports Medicine*, *26*(1), 64–74. https://doi.org/10.1080/15438627.2017.1393756

Camarillo, D. B., Shull, P. B., Mattson, J., Shultz, R., & Garza, D. (2013). An Instrumented Mouthguard for Measuring Linear and AngularHead Impact

Kinematics in American Football. *Annals of Biomedical Engineering2*, *41*(9), 1939–1949. https://doi.org/10.1007/s10439-013-0801-y

Campbell, K. R., Marshall, S. W., LUCK, J. F., PINTON, G. F., STITZEL, J. D., BOONE, J. S., GUSKIEWICZ, K. M., & MIHALIK, J. P. (2020). Head Impact Telemetry System's Video-based Impact Detection and Location Accuracy. *Medicine & Science in Sports & Exercise*, *52*(10), 2198–2206. https://doi.org/10.1249/MSS.0000000000002371

Capizzi, A., Woo, J., & Verduzco-Gutierrez, M. (2020). Traumatic Brain Injury. *Medical Clinics of North America*, *104*(2), 213–238. https://doi.org/10.1016/j.mcna.2019.11.001

Carroll, L., Cassidy, J. D., Peloso, P., Borg, J., von Holst, H., Holm, L., Paniak, C., & Pépin, M. (2004). Prognosis for mild traumatic brain injury: results of the who collaborating centre task force on mild traumatic brain injury. *Journal of Rehabilitation Medicine*, *36*(0), 84–105. https://doi.org/10.1080/16501960410023859

Chambers, R. M., Gabbett, T. J., Gupta, R., Josman, C., Bown, R., Stridgeon, P., & Cole, M. H. (2019a). Automatic detection of one-on-one tackles and ruck events using microtechnology in rugby union. *Journal of Science and Medicine in Sport*, *22*(7), 827–832. https://doi.org/10.1016/j.jsams.2019.01.001

Chambers, R. M., Gabbett, T. J., Gupta, R., Josman, C., Bown, R., Stridgeon, P., & Cole, M. H. (2019b). Automatic detection of one-on-one tackles and ruck events using microtechnology in rugby union. *Journal of Science and Medicine in Sport*, *22*(7), 827–832. https://doi.org/10.1016/j.jsams.2019.01.001

Changa, A. R., Vietrogoski, R. A., & Carmel, P. W. (2018). Dr Harrison Martland and the history of punch drunk syndrome. *Brain: A Journal of Neurology*, *141*(1), 318–321. https://doi.org/10.1093/brain/awx349

Chang, Y., Kim, N., & Stenfelt, S. (2018). Simulation of the power transmission of bone-conducted sound in a finite-element model of the human head. *Biomechanics and Modeling in Mechanobiology*, *17*(6), 1741–1755. https://doi.org/10.1007/s10237-018-1053-4

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, X. W., & Lin, X. (2014). Big Data Deep Learning: Challenges and Perspectives. *IEEE Access*, *2*, 514–525. https://doi.org/10.1109/ACCESS.2014.2325029

Cohen, A., & Kovacevic, J. (1996). Wavelets: the mathematical background. *Proceedings of the IEEE*, *84*(4), 514–522. https://doi.org/10.1109/5.488697

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms, Third Edition* (3rd ed.). The MIT Press.

Corsellis, J. A. N., Bruton, C. J., & Freeman-Browne, D. (1973). The Aftermath of Boxing. *Psychologial Medicine*, *3*, 270–303. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L20173271

Cortes, N., Lincoln, A. E., Myer, G. D., Hepburn, L., Higgins, M., Putukian, M., & Caswell, S. V. (2017). Video Analysis Verification of Head Impact Events Measured by Wearable Sensors. *American Journal of Sports Medicine*, *45*(10), 2379–2387. https://doi.org/10.1177/0363546517706703

Cosgrave, M., & Williams, S. (2019). The epidemiology of concussion in professional rugby union in Ireland. *Physical Therapy in Sport*, *35*, 99–105. https://doi.org/10.1016/j.ptsp.2018.11.010

Csizmadia, G., Liszkai-Peres, K., Ferdinandy, B., Miklósi, Á., & Konok, V. (2022). Human activity recognition of children with wearable devices using LightGBM machine learning. *Scientific Reports*, *12*(1), 5472. https://doi.org/10.1038/s41598-022-09521-1

Daneshvar, D. H., Nair, E. S., Baucom, Z. H., Rasch, A., Abdolmohammadi, B., Uretsky, M., Saltiel, N., Shah, A., Jarnagin, J., Baugh, C. M., Martin, B. M., Palmisano, J. N., Cherry, J. D., Alvarez, V. E., Huber, B. R., Weuve, J., Nowinski, C. J., Cantu, R. C., Zafonte, R. D., … Mez, J. (2023). Leveraging football accelerometer data to quantify associations between repetitive head

impacts and chronic traumatic encephalopathy in males. *Nature Communications 2023 14:1*, *14*(1), 1–14. https://doi.org/10.1038/s41467-023-39183-0

Daneshvar, D. H., Riley, D. O., Nowinski, C. J., McKee, A. C., Stern, R. A., & Cantu, R. C. (2011). Long-Term Consequences: Effects on Normal Development Profile After Concussion. *Physical Medicine and Rehabilitation Clinics of North America*, *22*(4), 683–700. https://doi.org/10.1016/j.pmr.2011.08.009

Dang, L. M., Min, K., Wang, H., Piran, J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, *108*, 107561. https://doi.org/10.1016/j.patcog.2020.107561

Davies, J. (2023, August 31). *Head injuries: Nearly 300 rugby players suing over brain damage*. Https://Www.Bbc.Co.Uk/News/Uk-Wales-66662278.

Den Hartog, J. P. (1985). *Mechanical Vibrations*. Courier Corporation.

Diversified Technical Systems. (2024, December 1). *Angular rate sensor (ARS): ARS Pro and ARS HG*. Https://Dtsweb.Com/Angular-Rate-Sensor-Ars-pro-Ars-Hg/.

Domel, A. G., Raymond, S. J., Giordano, C., Liu, Y., Yousefsani, S. A., Fanton, M., Cecchi, N. J., Vovk, O., Pirozzi, I., Kight, A., Avery, B., Boumis, A., Fetters, T., Jandu, S., Mehring, W. M., Monga, S., Mouchawar, N., Rangel, I., Rice, E., … Camarillo, D. B. (2021). A new open-access platform for measuring and sharing mTBI data. *Scientific Reports*, *11*(1), 7501. https://doi.org/10.1038/s41598-021-87085-2

Dompier, T. P., Kerr, Z. Y., Marshall, S. W., Hainline, B., Snook, E. M., Hayden, R., & Simon, J. E. (2015). Incidence of Concussion During Practice and Games in Youth, High School, and Collegiate American Football Players. *JAMA Pediatrics*, *169*(7), 659. https://doi.org/10.1001/jamapediatrics.2015.0210

Fraas, M. R., Coughlan, G. F., Hart, E. C., & McCarthy, C. (2014). Concussion history and reporting rates in elite Irish rugby union players. *Physical Therapy in Sport*, *15*(3), 136–142. https://doi.org/10.1016/j.ptsp.2013.08.002

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Fujita, M., Wei, E. P., & Povlishock, J. T. (2012). Intensity-and interval-specific repetitive traumatic brain injury can evoke both axonal and microvascular damage. *Journal of Neurotrauma*, *29*(12), 2172–2180. https://doi.org/10.1089/neu.2012.2357

Fuller, C. W., Laborde, F., Leather, R. J., & Molloy, M. G. (2008). International Rugby Board Rugby World Cup 2007 injury surveillance study. *British Journal of Sports Medicine*, *42*(6), 452–459. https://doi.org/10.1136/bjsm.2008.047035

Fuller, C. W., Laborde, F., Leather, R. J., & Molloy, M. G. (2016). Rugby World Cup 2015: World Rugby injury surveillance study. *British Journal of Sports Medicine*, *51*(1), 51–57. https://doi.org/10.1136/bjsports-2016-096275

Fuller, C. W., Sheerin, K., & Targett, S. (2013). Rugby World Cup 2011: International Rugby Board Injury Surveillance Study. *British Journal of Sports Medicine*, *42*(18), 452–459. https://doi.org/10.1136/bjsports-2012-091155

Fuller, C. W., Taylor, A., & Douglas, M. (2020). Rugby World Cup 2019 injury surveillance study. *South African Journal of Sports Medicine*, *32*(1), 1–6. https://doi.org/10.17159/2078-516X/2020/v32i1a8062

Gabler, L., Huddlestone, S., Dau, N., Lessley, D., Arbogast, K. B., Thompson, X., Resch, J., & Crandall, J. (2020). On-Field Performance of an Instrumented Mouthguard for Detecting Head Impacts in American Football. *Annals of Biomedical Engineering*, *48*(11), 2599–2612. https://doi.org/https://doi.org/10.1007/s10439-020-02654-2

Gabler, L., Patton, D., Begonia, M., Daniel, R., Rezaei, A., Huber, C., Siegmund, G., Rooks, T., & Wu, L. (2022). Consensus Head Acceleration Measurement Practices (CHAMP): Laboratory Validation of Wearable Head Kinematic Devices. *Annals of Biomedical Engineering*. https://doi.org/10.1007/s10439-022-03066-0

García, S., Luengo, J., Herrera, F., García, S., Luengo, J., & Herrera, F. (2015). Feature selection. *Intelligent Systems Reference Library*, *72*(6), 163–193. https://doi.org/10.1007/978-3-319-10247-4_7

Gardner, R. C., & Yaffe, K. (2015). Epidemiology of mild traumatic brain injury and neurodegenerative disease. *Molecular and Cellular Neuroscience*, *66*, 75–80. https://doi.org/10.1016/j.mcn.2015.03.001

Gastin, P., McLean, O., Breed, R., & Spittle, M. (2014). Tackle and impact detection in elite Australian football using wearable microsensor technology. *Journal of Sports Sciences*, *32*(10).

Gavett, B., Stern, R., & McKee, A. (2011). Chronic Traumatic Encephalopathy: A Potential Late Effect of Sport-Related Concussive and Subconcussive Head Trauma. *Clinical Journal of Sport Medicine2*, *30*(1), 1–10. https://doi.org/10.1016/j.csm.2010.09.007.Chronic

Gellner, R., Begonia, M., & Rowson, S. (2024). Choosing Optimal Cutoff Frequencies for Filtering Linear Acceleration and Angular Velocity Signals Associated with Head Impacts Measured by Instrumented Mouthguards. *Annals of Biomedical Engineering*, *52*(5), 1415–1424. https://doi.org/10.1007/s10439-024-03466-4

Gennarelli, T. A., Thibault, L. E., Tomei, G., Wiser, R., Graham, D., & Adams, J. (1987). Directional dependence of axonal brain injury due to centroidal and non-centroidal acceleration. *SAE Technical Papers*, *1*. https://doi.org/10.4271/872197

Giza, C. C., & Hovda, D. A. (2001). The Neurometabolic Cascade of Concussion. *Journal of Athletic Training*, *36*(3), 228–235.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Goodin, P., Gardner, A. J., Dokani, N., Nizette, B., Ahmadizadeh, S., Edwards, S., & Iverson, G. L. (2021). Development of a Machine-Learning-Based Classifier for the Identification of Head and Body Impacts in Elite Level Australian Rules Football Players. *Frontiers in Sports and Active Living*, *3*. https://doi.org/10.3389/fspor.2021.725245

Green, J. I. (2017). The Role of Mouthguards in Preventing and Reducing Sports-Related Trauma. *Primary Dental Journal*, *6*(2), 27–34. https://doi.org/10.1308/205016817821281738

Greybe, D. G., Jones, C. M., Brown, M. R., & Williams, E. M. P. (2020). Comparison of head impact measurements via an instrumented mouthguard and an

anthropometric testing device. *Sports Engineering*, *23*(1), 1–11. https://doi.org/10.1007/s12283-020-00324-z

Gurdjian, E. S., Roberts, V. L., & Thomas, L. M. (1966). Tolerance curves of acceleration and intracranial pressure and protective index in experimental head injury. *Journal of Trauma*, *6*(5), 600–604. https://doi.org/10.1097/00005373-196609000-00005

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, C. R., Millman, K. J., Walt, S. J. Van Der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., & Kerkwijk, M. H. Van. (2020). Array programming with NumPy. *Nature*, *585*(June), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). *Overview of Supervised Learning* (pp. 9–41). https://doi.org/10.1007/978-0-387-84858-7_2

Hastie, T., Tibshirani, R., & Friedman, J. (2009b). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Hecht-Nielsen. (1989). Theory of the backpropagation neural network. *International Joint Conference on Neural Networks*, 593–605 vol.1. https://doi.org/10.1109/IJCNN.1989.118638

Hedin, D. S., Gibson, P. L., Bartsch, A. J., & Samorezov, S. (2016). Development of a head impact monitoring “Intelligent Mouthguard”. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2007–2009. https://doi.org/10.1109/EMBC.2016.7591119

Hendry, D., Chai, K., Campbell, A., Hopper, L., O'Sullivan, P., & Straker, L. (2020). Development of a Human Activity Recognition System for Ballet Tasks. *Sports Medicine - Open*, *6*(1), 10. https://doi.org/10.1186/s40798-020-0237-5

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and Vision Computing*, *60*, 4–21. https://doi.org/10.1016/j.imavis.2017.01.010

Horng, D., Chau, Tech, G., Roozbahani, M., Heer, J., Stasko, J., & Faloutsos, C. (2019). Ensemble Methods. *Machine Learning with Spark™ and Python®*, 449–523. https://doi.org/10.1017/9781139236003.012

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley. https://doi.org/10.1002/9781118548387

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics (Oxford, England)*, *21*(8), 1509–1515. https://doi.org/10.1093/bioinformatics/bti171

Ingle, S. (2023, November 30). Rugby players to apply for class action lawsuit in legal case over brain injuries. *The Guardian*.

Jadischke, R., Viano, D. C., Dau, N., King, A. I., & McCarthy, J. (2013). On the accuracy of the head impact telemetry (hit) system used in football helmets. *Journal of Biomechanics*, *46*(13), 2310–2315. https://doi.org/10.1016/j.jbiomech.2013.05.030

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jobanputra, C., Bavishi, J., & Doshi, N. (2019). Human activity recognition: A survey. *Procedia Computer Science*, *155*(2018), 698–703. https://doi.org/10.1016/j.procs.2019.08.100

Jones, B., Tooby, J., Weaving, D., Till, K., Owen, C., Begonia, M., Stokes, K. A., Rowson, S., Phillips, G., Hendricks, S., Falvey, É. C., Al-Dawoud, M., & Tierney, G. (2022a). Ready for impact? A validity and feasibility study of

instrumented mouthguards (iMGs). *British Journal of Sports Medicine*. https://doi.org/10.1136/bjsports-2022-105523

Jones, B., Tooby, J., Weaving, D., Till, K., Owen, C., Begonia, M., Stokes, K. A., Rowson, S., Phillips, G., Hendricks, S., Falvey, É. C., Al-Dawoud, M., & Tierney, G. (2022b). Ready for impact? A validity and feasibility study of instrumented mouthguards (iMGs). *British Journal of Sports Medicine*, *56*(20), 1171–1179. https://doi.org/10.1136/bjsports-2022-105523

Kautz, T., Groh, B., Hannick, J., Jensen, U., Strubberg, H., & Eskofier, B. (2017). Activity recognition in beach volleyball using a Deep Convolutional Neural Network: Leveraging the potential of Deep Learning in sports. *Data Mining and Knowledge Discovery*, *31*(6), 1678–1705. https://doi.org/10.1007/s10618-017-0495-0

Kemp, S., Stokes, K., Mckay, C., & Roberts, S. (2022). *BUCS Super Rugby Injury Surveillance Project 2019-20*.

Khusainov, R., Azzi, D., Achumba, I., & Bersch, S. (2013). Real-Time Human Ambulation, Activity, and Physiological Monitoring: Taxonomy of Issues, Techniques, Applications, Challenges and Limitations. *Sensors*, *13*(10), 12852–12902. https://doi.org/10.3390/s131012852

Kieffer, E. E., Begonia, M. T., Tyson, A. M., & Rowson, S. (2020). A Two-Phased Approach to Quantifying Head Impact Sensor Accuracy: In-Laboratory and On-Field Assessments. *Annals of Biomedical Engineering*, *48*(11), 2613–2625. https://doi.org/10.1007/s10439-020-02647-1

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out". *Health Information Management Journal*, *47*(3), 103–105. https://doi.org/10.1177/1833358318774357

King, D. A., Hume, P. A., Gissane, C., Kieser, D. C., & Clark, T. N. (2018). Head impact exposure from match participation in women's rugby league over one season of domestic competition. *Journal of Science and Medicine in Sport*, *21*(2), 139–146. https://doi.org/10.1016/j.jsams.2017.10.026

King, D. A., Hume, P. A., Hind, K., Clark, T. N., & Hardaker, N. (2022). The Incidence, Cost, and Burden of Concussion in Women's Rugby League and Rugby Union: A Systematic Review and Pooled Analysis. *Sports Medicine*, *52*(8), 1751–1764. https://doi.org/10.1007/s40279-022-01645-8

King, D. A., Hume, P., Gissane, C., Brughelli, M., & Clark, T. (2016). The Influence of Head Impact Threshold for Reporting Data in Contact and Collision Sports: Systematic Review and Original Data Analysis. *Sports Medicine*, *46*(2), 151–169. https://doi.org/10.1007/s40279-015-0423-7

Kingsford, C., & Salzberg, S. (2008). What are decision trees? *Natural Biotechnology*, *26*(9), 1011–1013. https://doi.org/10.1038/nbt0908-1011.What

Knapik, J. J., Marshall, S. W., Lee, R. B., Darakjy, S. S., Jones, S. B., Mitchener, T. A., delaCruz, G. G., & Jones, B. H. (2007). Mouthguards in Sport Activities. *Sports Medicine*, *37*(2), 117–144. https://doi.org/10.2165/00007256-200737020-00003

Koerte, I. K., Esopenko, C., Hinds, S. R., Shenton, M. E., Bonke, E. M., Bazarian, J. J., Bickart, K. C., Bigler, E. D., Bouix, S., Buckley, T. A., Choe, M. C., Echlin, P. S., Gill, J., Giza, C. C., Hayes, J., Hodges, C. B., Irimia, A., Johnson, P. K., Kenney, K., … Baron, D. (2021). The ENIGMA sports injury working group:– an international collaboration to further our understanding of sport-related brain injury. *Brain Imaging and Behavior*, *15*(2), 576–584. https://doi.org/10.1007/s11682-020-00370-y

Komarek, P., & Moore, A. (2003). Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs. *International Conference on Artificial Intelligence and Statistics*.

Kong, Y., & Fu, Y. (2022). Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, *130*(5), 1366–1401. https://doi.org/10.1007/S11263-022-01594-9

Kuo, C., Wu, L. C., Hammoor, B., Luck, J., Cutcliffe, H., Lynall, R., Kait, J., Campell, K., Mihalik, J., Bass, C., & Camarillo, D. B. (2016). Effect of the mandible on mouthguard measurements of head kinematics. *Journal of Biomechanics*, *49*(9), 1845–1853. https://doi.org/10.1016/j.jbiomech.2016.04.017

Kuo, C., Wu, L. C., Zhao, W., Fanton, M., Ji, S., & Camarillo, D. B. (2016). Propagation of Errors from Skull Kinematic Measurements to Finite Element Tissue Responses. *Biomechanics and Modeling in Mechanobiology*, *176*(1), 139–148. https://doi.org/10.1007/s10237-017-0957-8

Kuo, C., Wu, L., Loza, J., Senif, D., Anderson, S. C., & Camarillo, D. B. (2017). Comparison of video-based and sensor-based head impact exposure. *BioRxiv*, 1–19. https://doi.org/10.1101/235432

Kusunoki, T., Matsuoka, J., Ohtsu, H., Kagimura, T., & Nakamura, H. (2009). Relationship between Intraclass and Concordance Correlation Coefficients: Similarities and Differences. *Japanese Journal of Biometrics*, *30*(1), 35–53. https://doi.org/10.5691/JJB.30.35

Lamond, L. C., Caccese, J. B., Buckley, T. A., Glutting, J., & Kaminski, T. W. (2018). Linear Acceleration in Direct Head Contact Across Impact Type, Player Position, and Playing Scenario in Collegiate Women's Soccer Players. *Journal of Athletic Training*, *53*(2), 115–121. https://doi.org/10.4085/1062-6050-90-17

Lara, O., & Labrador, M. (2013). Human activity recognition using wearable sensors. *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, *15*(3), 1192–1209. https://doi.org/10.1007/978-981-15-5679-1_51

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K., & O'Leary, A. (2019). PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, *4*(36), 1237. https://doi.org/10.21105/joss.01237

Lee, T. H., Ullah, A., & Wang, R. (2020). Bootstrap Aggregating and Random Forest. *Advanced Studies in Theoretical and Applied Econometrics*, *52*, 389–429. https://doi.org/10.1007/978-3-030-31150-6_13

Le Flao, E., Siegmund, G. P., & Borotkanics, R. (2022). Head Impact Research Using Inertial Sensors in Sport: A Systematic Review of Methods, Demographics, and Factors Contributing to Exposure. *Sports Medicine*, *52*(3), 481–504. https://doi.org/10.1007/s40279-021-01574-y

Liaw, A., & Wiener, M. (2001). Classification and Regression by RandomForest. *Forest*, *23*.

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, *45*(1), 255. https://doi.org/10.2307/2532051

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008a). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008b). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(1). https://doi.org/10.1145/2133360.2133363

Liu, Y., Domel, A. G., Yousefsani, S. A., Kondic, J., Grant, G., Zeineh, M., & Camarillo, D. B. (2020). Validation and Comparison of Instrumented Mouthguards for Measuring Head Kinematics and Assessing Brain Deformation in Football Impacts. *Annals of Biomedical Engineering*, *48*(11), 2580–2598. https://doi.org/10.1007/s10439-020-02629-3

Liu, Y., Nie, L., Liu, L., & Rosenblum, D. S. (2016). From action to activity: Sensor-based activity recognition. *Neurocomputing*, *181*, 108–115. https://doi.org/10.1016/j.neucom.2015.08.096

Loane, D. J., & Faden, A. I. (2010). Neuroprotection for traumatic brain injury: translational challenges and emerging therapeutic strategies. *Trends in Pharmacological Sciences*, *31*(12), 596–604. https://doi.org/10.1016/j.tips.2010.09.005

Lonini, L., Dai, A., Shawen, N., Simuni, T., Poon, C., Shimanovich, L., Daeschler, M., Ghaffari, R., Rogers, J. A., & Jayaraman, A. (2018). Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0071-z

Luke, D., Kenny, R., Bondi, D., Clansey, A. C., & Wu, L. C. (2024). On-field instrumented mouthguard coupling. *Journal of Biomechanics*, *162*, 111889. https://doi.org/10.1016/j.jbiomech.2023.111889

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lutz, M. (2013). *Learning python: Powerful object-oriented programming.* O'Reilly Media, Inc.

Maas, A. I. R., Menon, D. K., Adelson, P. D., Andelic, N., Bell, M. J., Belli, A., Bragge, P., Brazinova, A., Büki, A., Chesnut, R. M., Citerio, G., Coburn, M., Cooper, D. J., Crowder, A. T., Czeiter, E., Czosnyka, M., Diaz-Arrastia, R., Dreier, J. P., Duhaime, A.-C., … Zumbo, F. (2017). Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, *16*(12), 987–1048. https://doi.org/10.1016/S1474-4422(17)30371-X

Mack, C. D., Solomon, G., Covassin, T., Theodore, N., Cárdenas, J., & Sills, A. (2021). Epidemiology of Concussion in the National Football League, 2015-2019. *Sports Health: A Multidisciplinary Approach*, *13*(5), 423–430. https://doi.org/10.1177/19417381211011446

Mahon, G. (2005). *A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient.*

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *WIREs Computational Statistics*, *1*(3), 283–289. https://doi.org/10.1002/wics.49

Martland, H. S. (1928). Punch Drunk. *Journal of the American Medical Association*, *91*(15), 1103–1107.

McCrory, P. R., Meeuwisse, W., Dvořák, J., Aubry, M., Bailes, J., Broglio, S., Cantu, R. C., Cassidy, D., Echemendia, R. J., Castellani, R. J., Davis, G. A., Ellenbogen,

R., Emery, C., Engebretsen, L., Feddermann-Demont, N., Giza, C. C., Guskiewicz, K. M., Herring, S., Iverson, G. L., … Vos, P. E. (2017). Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. *British Journal of Sports Medicine*, *51*(11), 838–847. https://doi.org/10.1136/bjsports-2017-097699

McKee, A. C. (2020). The Neuropathology of Chronic Traumatic Encephalopathy: The Status of the Literature. *Seminars in Neurology*, *40*(04), 359–369. https://doi.org/10.1055/s-0040-1713632

McKee, A. C., Cantu, R. C., Nowinski, C. J., Hedley-Whyte, E. T., Gavett, B. E., Budson, A. E., Santini, V. E., Lee, H.-S., Kubilus, C. A., & Stern, R. A. (2009). Chronic Traumatic Encephalopathy in Athletes: Progressive Tauopathy After Repetitive Head Injury. *Journal of Neuropathology & Experimental Neurology*, *68*(7), 709–735. https://doi.org/10.1097/NEN.0b013e3181a9d503

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *SciPy*, 56–61. https://doi.org/10.25080/MAJORA-92BF1922-00A

McKinney, W. (2011). *pandas: a Foundational Python Library for Data Analysis and Statistics*.

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media, Inc.

McNamee, M., Anderson, L. C., Borry, P., Camporesi, S., Derman, W., Holm, S., Knox, T. R., Leuridan, B., Loland, S., Lopez Frias, F. J., Lorusso, L., Malcolm, D., McArdle, D., Partridge, B., Schramme, T., & Weed, M. (2023). Sport-related concussion research agenda beyond medical science: culture, ethics, science, policy. *Journal of Medical Ethics*, jme-2022-108812. https://doi.org/10.1136/jme-2022-108812

Menard, S. (2002). *Applied Logistic Regression Analysis*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983433

Menon, D. K., Schwab, K., Wright, D. W., & Maas, A. I. (2010). Position Statement: Definition of Traumatic Brain Injury. *Archives of Physical Medicine and Rehabilitation*, *91*(11), 1637–1640. https://doi.org/10.1016/j.apmr.2010.05.017

Meythaler, J. M., Peduzzi, J. D., Eleftheriou, E., & Novack, T. A. (2001). Current concepts: Diffuse axonal injury–associated traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *82*(10), 1461–1471. https://doi.org/10.1053/apmr.2001.25137

Mez, J., Solomon, T., Daneshvar, D., Stein, T., & McKee, A. (2016). Pathologically Confirmed Chronic Traumatic Encephalopathy in a 25-Year-Old Former College Football Player. *JAMAM*, *176*(1), 139–148. https://doi.org/10.1001/jamaneurol.2015.3998.Pathologically

M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Miller, L. E., Kuo, C., Wu, L. C., Urban, J. E., Camarillo, D. B., & Stitzel, J. D. (2018). Validation of a Custom Instrumented Retainer Form Factor for Measuring Linear and Angular Head Impact Kinematics. *Journal of Biomechanical Engineering*, *140*(5). https://doi.org/10.1115/1.4039165

Milspaugh, J. A. (1937). Dementia Pugilistica. *US Naval Medical Bulletin*.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2023). Explainable Artificial Intelligence (XAI). *International Journal of Food and Nutritional Science*, *55*(5), 3503–3568. https://doi.org/10.1007/S10462-021-10088-Y

Mitchell, T. M. (1997). *Machine Learning* (1st ed.). McGraw-Hill. https://doi.org/https://doi.org/10.1002/(SICI)1099-1689(199909)9:3<191::AID-STVR184>3.0.CO;2-E

Moguerza, J. M., & Muñoz, A. (2006). Support Vector Machines with Applications. *Statistical Science*, *21*(3), 322–336. https://doi.org/10.1214/088342306000000493

Mohan, M., & Huynh, L. (2019). Sex Differences in the Spine. *Current Physical Medicine and Rehabilitation Reports*, *7*(3), 246–252. https://doi.org/10.1007/s40141-019-00234-7

Mohan, M., Weaving, D., Gardner, A. J., Hendricks, S., Stokes, K., Phillips, G., Cross, M., & Jones, P. Ben. (2024). 776 BO01 – Can a novel computer vision-based

framework detect head-on-head impacts during a rugby league tackle? *Brief Oral Abstracts*, A43.2-A44. https://doi.org/10.1136/BJSPORTS-2024-IOC.77

Montenigro, P. H., Alosco, M. L., Martin, B. M., Daneshvar, D. H., Mez, J., Chaisson, C. E., Nowinski, C. J., Au, R., McKee, A. C., Cantu, R. C., McClean, M. D., Stern, R. A., & Tripodis, Y. (2017). Cumulative Head Impact Exposure Predicts Later-Life Depression, Apathy, Executive Dysfunction, and Cognitive Impairment in Former High School and College Football Players. *Journal of Neurotrauma*, *34*(2), 328–340. https://doi.org/10.1089/NEU.2016.4413/ASSET/IMAGES/LARGE/FIGURE2.JPEG

Montenigro, P. H., Corp, D. T., Stein, T. D., Cantu, R. C., & Stern, R. A. (2015). Chronic Traumatic Encephalopathy: Historical Origins and Current Perspective. *Annual Review of Clinical Psychology*, *11*(1), 309–330. https://doi.org/10.1146/annurev-clinpsy-032814-112814

Moon, D. W., Beedle, C. W., & Kovacic, C. R. (1971). Peak head acceleration of athletes during competition–football. *Medicine and Science in Sports*, *3*(1), 44–50.

Moore, I. S., Ranson, C., & Mathema, P. (2015). Injury Risk in International Rugby Union: Three-Year Injury Surveillance of the Welsh National Team. *Orthopaedic Journal of Sports Medicine*, *3*(7), 1–9. https://doi.org/10.1177/2325967115596194

Murphy, K. P. (2012). Machine learning - a probabilistic perspective. *Adaptive Computation and Machine Learning Series*.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5–6), 183–197. https://doi.org/10.1016/0925-2312(91)90023-5

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Ng, S. Y., & Lee, A. Y. W. (2019). Traumatic Brain Injuries: Pathophysiology and Potential Therapeutic Targets. *Frontiers in Cellular Neuroscience*, *13*. https://doi.org/10.3389/fncel.2019.00528

Ng, T., Bussone, W., & Duma, S. (2006). The effect of gender and body size on linear accelerations of the head observed during daily activities. *Biomedical Scientific Instrumentation*, *42*, 25–30.

Noble, W. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Nowinski, C. J., Bureau, S. C., Buckland, M. E., Curtis, M. A., Daneshvar, D. H., Faull, R. L. M., Grinberg, L. T., Hill-Yardin, E. L., Murray, H. C., Pearce, A. J., Suter, C. M., White, A. J., Finkel, A. M., & Cantu, R. C. (2022). Applying the Bradford Hill Criteria for Causation to Repetitive Head Impacts and Chronic Traumatic Encephalopathy. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.938163

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, *9*(3), 10–20. https://doi.org/10.1109/MCSE.2007.58

Omalu, B. I., DeKosky, S. T., Minster, R. L., Kamboh, M. I., Hamilton, R. L., & Wecht, C. H. (2005). Chronic traumatic encephalopathy in a National Football League player. *Neurosurgery*, *57*(1), 128–133. https://doi.org/10.1227/01.NEU.0000163407.92769.ED

Palmer, A., & Hargreaves, Z. (2023, November 29). *Concussion: Research in women's game lacking - World Rugby*. Https://Www.Bbc.Co.Uk/News/Uk-Wales-67554363.

Pareek, P., & Thakkar, A. (2021). A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259–2322. https://doi.org/10.1007/s10462-020-09904-8

Paris, A. J., Antonini, K. R., & McFerran Brock, J. (2010). Accelerations of the head during soccer ball heading. *Summer Bioengineering Conference*, *44038*, 815–816.

Patricios, J. S., Ardern, C. L., Hislop, M. D., Aubry, M., Bloomfield, P., Broderick, C., Clifton, P., Echemendia, R. J., Ellenbogen, R. G., Falvey, É. C., Fuller, G. W., Grand, J., Hack, D., Harcourt, P. R., Hughes, D., McGuirk, N., Meeuwisse, W., Miller, J., Parsons, J. T., … Raftery, M. (2018). Implementation of the 2017 Berlin Concussion in Sport Group Consensus Statement in contact and collision sports: a joint position statement from 11 national and international sports organisations. *British Journal of Sports Medicine*, *52*(10), 635–641. https://doi.org/10.1136/bjsports-2018-099079

Patricios, J. S., Schneider, K. J., Dvorak, J., Ahmed, O. H., Blauwet, C., Cantu, R. C., Davis, G. A., Echemendia, R. J., Makdissi, M., McNamee, M., Broglio, S., Emery, C. A., Feddermann-Demont, N., Fuller, G. W., Giza, C. C., Guskiewicz, K. M., Hainline, B., Iverson, G. L., Kutcher, J. S., … Meeuwisse, W. (2023). Consensus statement on concussion in sport: the 6th International Conference on Concussion in Sport–Amsterdam, October 2022. *British Journal of Sports Medicine*, *57*(11), 695–711. https://doi.org/10.1136/bjsports-2023-106898

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *IARJSET*, 20–22. https://doi.org/10.17148/IARJSET.2015.2305

Patton, D. A. (2016). A Review of Instrumented Equipment to Investigate Head Impacts in Sport. *Applied Bionics and Biomechanics*, *2016*. https://doi.org/10.1155/2016/7049743

Patton, D. A., Huber, C. M., Jain, D., Myers, R. K., McDonald, C. C., Margulies, S. S., Master, C. L., & Arbogast, K. B. (2020). Head Impact Sensor Studies In Sports: A Systematic Review Of Exposure Confirmation Methods. *Annals of Biomedical Engineering*, *48*(11), 2497–2507. https://doi.org/10.1007/s10439-020-02642-6

Patton, D. A., Huber, C. M., McDonald, C. C., Margulies, S. S., Master, C. L., & Arbogast, K. B. (2020). Video Confirmation of Head Impact Sensor Data From High School Soccer Players. *The American Journal of Sports Medicine*, *48*(5), 1246–1253. https://doi.org/10.1177/0363546520906406

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Peregino, P., & Bukowski, E. (2004). *Development and Evaluation of a Surface-Mount, High-G Accelerometer*.

Pervez, M., Kitagawa, R. S., & Chang, T. R. (2018). Definition of Traumatic Brain Injury, Neurosurgery, Trauma Orthopedics, Neuroimaging, Psychology, and Psychiatry in Mild Traumatic Brain Injury. *Neuroimaging Clinics of North America*, *28*(1), 1–13. https://doi.org/10.1016/j.nic.2017.09.010

Pfaller, A. Y., Brooks, M. A., Hetzel, S., & McGuine, T. A. (2019). Effect of a New Rule Limiting Full Contact Practice on the Incidence of Sport-Related Concussion in High School Football Players. *The American Journal of Sports Medicine*, *47*(10), 2294–2299. https://doi.org/10.1177/0363546519860120

Pfister, B. J., Chickola, L., & Smith, D. H. (2009). Head motions while riding roller coasters: Implications for brain injury. *American Journal of Forensic Medicine and Pathology*, *30*(4), 339–345. https://doi.org/10.1097/PAF.0B013E318187E0C9

Phillips, D. (2010). *Python 3 Object Oriented Programming*.

Pisner, D., & Schnyer, D. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/B9780128157398000067

Powell, D. R. L., Petrie, F. J., Docherty, P. D., Arora, H., & Williams, E. M. P. (2023). Development of a Head Acceleration Event Classification Algorithm for Female Rugby Union. *Annals of Biomedical Engineering*. https://doi.org/10.1007/s10439-023-03138-9

Press, J. N., & Rowson, S. (2017). Quantifying head impact exposure in collegiate women's soccer. *Clinical Journal of Sport Medicine*, *27*(2), 104–110. https://doi.org/10.1097/JSM.0000000000000313

Raftery, M., & Falvey, É. C. (2022). Rugby's implementation lessons: the importance of a 'compliance wedge' to support successful implementation for injury

prevention. *British Journal of Sports Medicine*, *56*(1), 1–2. https://doi.org/10.1136/bjsports-2020-103454

Raftery, M., Tucker, R., & Falvey, É. C. (2021). Getting tough on concussion: how welfare-driven law change may improve player safety—a Rugby Union experience. *British Journal of Sports Medicine*, *55*(10), 527–529. https://doi.org/10.1136/bjsports-2019-101885

Rane, N., Choudhary, S., & Rane, J. (2024). Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4849885

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*.

Raymond, S. J., Cecchi, N. J., Alizadeh, H. V., Callan, A. A., Rice, E., Liu, Y., Zhou, Z., Zeineh, M., & Camarillo, D. B. (2022). Physics-Informed Machine Learning Improves Detection of Head Impacts. *Annals of Biomedical Engineering*. https://doi.org/10.1007/s10439-022-02911-6

Reid, S. E., Tarkington, J., Epstein, H. M., & O'Dea, T. J. (1971). Brain tolerance to impact in football. *Surgery, Gynecology & Obstetrics*, *133 6*, 929–936. https://api.semanticscholar.org/CorpusID:6891460

Rezaei, A., & Wu, L. C. (2022). Automated soccer head impact exposure tracking using video and deep learning. *Scientific Reports*, *12*(1). https://doi.org/10.1038/S41598-022-13220-2

Rioul, O., & Duhamel, P. (1992). Fast algorithms for discrete and continuous wavelet transforms. *IEEE Transactions on Information Theory*, *38*(2), 569–586. https://doi.org/10.1109/18.119724

Roberts, S., Mckay, C., Stokes, K., & Kemp, S. (2023). *Community Rugby Injury Surveillance and Prevention Project 2021-22*. https://keepyourbootson.co.uk/rugbysafe-toolkit/research/

Rokach, L. (2019). *Ensemble Learning: Pattern Classification Using Ensemble Methods (Second Edition)* (2nd ed.). World Scientific Publishing Co Pte Ltd.

Rokach, L., & Maimon, O. (2005). Decision Trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_9

Roozenbeek, B., Maas, A. I. R., & Menon, D. K. (2013). Changing patterns in the epidemiology of traumatic brain injury. *Nature Reviews Neurology*, *9*(4), 231–236. https://doi.org/10.1038/nrneurol.2013.22

Rossum, G. (2000). *Python Reference Manual*.

Rossum, G. (2007). Python Programming Language. *USENIX Annual Technical Conference*.

Rossum, G., & Drake, F. (2009). *Python 3 Reference Manual*.

Rowlands, A. V., & Stiles, V. H. (2012). Accelerometer counts and raw acceleration output in relation to mechanical loading. *Journal of Biomechanics*, *45*(3), 448–454. https://doi.org/10.1016/j.jbiomech.2011.12.006

Roy, P. K., & Om, H. (2018). Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos. *Studies in Computational Intelligence*, *730*, 277–294. https://doi.org/10.1007/978-3-319-63754-9_13

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rutherford, A., Stewart, W., & Bruno, D. (2019). Heading for trouble: is dementia a game changer for football? *British Journal of Sports Medicine*, *53*(6), 321–322. https://doi.org/10.1136/bjsports-2017-097627

Ryan, H. (1994). Ricker, Ormsby, Klauder, Butterworth - A Choice of Wavelets . *Canadian Society of Exploration Geophysicists*, *19*(07), 8–9.

Saatman, K. E., Duhaime, A.-C., Bullock, R., Maas, A. I. R., Valadka, A., & Manley, G. T. (2008). Classification of Traumatic Brain Injury for Targeted Therapies. *Journal of Neurotrauma*, *25*(7), 719–738. https://doi.org/10.1089/neu.2008.0586

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/WIDM.1249

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salman, S., & Liu, X. (2019). Overfitting Mechanism and Avoidance in Deep Neural Networks. *ArXiv.Org*.

Salmon, D., Sullivan, J., Romanchuk, J., Murphy, I., Walters, S., Whatman, C., Clacy, A., Keung, S., & Van Der Vis, K. (2019). Infographic. New Zealand rugby's community concussion initiative: keeping kiwi communities RugbySmart. *British Journal of Sports Medicine*, *54*(5), 300–301. https://doi.org/10.1136/BJSPORTS-2019-100949

Sands, W. A., Kelly, B., Bogdanis, G., Barker, L., Donti, O., McNeal, J. R., & Penitente, G. (2019). COMPARISON OF BUNGEE-AIDED AND FREE-BOUNCING ACCELERATIONS ON TRAMPOLINE. *Science of Gymnastics Journal*, *11*(3), 279–288. https://doi.org/10.52165/SGJ.11.3.279-288

Sanner, M. F. (1999). Python: a programming language for software integration and development. *Journal of Molecular Graphics & Modelling*, *17*(1), 57–61.

Schneider, R. C. (1961). Serious and Fatal Football Injuries Involving the Head and Spinal Cord. *JAMA*, *177*(6), 362. https://doi.org/10.1001/jama.1961.03040320006002

Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ*, *345*(jul04 1), e4483–e4483. https://doi.org/10.1136/bmj.e4483

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

Sharma, S., & Sharma, V. (2016). Performance of Various Machine Learning Classifiers on Small Datasets with Varying Dimensionalities: A Study.

Circulation in Computer Science, *1*(1), 30–35. https://doi.org/10.22632/ccs-2016-251-23

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Silva, V. S. da, & Vieira, M. F. S. (2020). International Society for the Advancement of Kinanthropometry (ISAK) Global: international accreditation scheme of the competent anthropometrist. *Revista Brasileira de Cineantropometria & Desempenho Humano*, *22*. https://doi.org/10.1590/1980-0037.2020v22e70517

Society of Automotive Engineers. (2003). *Instrumentation for Impact Test-Part 1-Electronic Instrumentation-SAE J211/1*.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Solomon Jr, O. (1991). *PSD computations using Welch's method. [Power Spectral Density (PSD)]*.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. https://doi.org/10.11613/BM.2014.003

Srinath, K. R. (2017). Python – The Fastest Growing Programming Language. *International Research Journal of Engineering and Technology (IRJET)*, *4*(12), 354–357.

Starling, L. T., Gabb, N., Williams, S., Kemp, S., & Stokes, K. A. (2023). Longitudinal study of six seasons of match injuries in elite female rugby union. *British Journal of Sports Medicine*, *57*(4), 212–217. https://doi.org/10.1136/BJSPORTS-2022-105831

Stein, T. D., Alvarez, V. E., & McKee, A. C. (2014). Chronic traumatic encephalopathy: a spectrum of neuropathological changes following repetitive brain trauma in athletes and military personnel. *Alzheimer's Research & Therapy*, *6*(1), 4. https://doi.org/10.1186/alzrt234

Stemper, B. D., Derosia, J. J., Yogananan, N., Pintar, F. A., Shender, B. S., & Paskoff, G. R. (2009). Gender dependent cervical spine anatomical differences in size-matched volunteers - biomed 2009. *Biomedical Sciences Instrumentation*, *45*, 149–154.

Stemper, B. D., Shah, A. S., Harezlak, J., Rowson, S., Duma, S., Mihalik, J. P., Riggen, L. D., Brooks, A., Cameron, K. L., Giza, C. C., Houston, M. N., Jackson, J., Posner, M. A., McGinty, G., DiFiori, J., Broglio, S. P., McAllister, T. W., & McCrea, M. (2019). Repetitive Head Impact Exposure in College Football Following an NCAA Rule Change to Eliminate Two-A-Day Preseason Practices: A Study from the NCAA-DoD CARE Consortium. *Annals of Biomedical Engineering*, *47*(10), 2073–2085. https://doi.org/10.1007/s10439-019-02335-9

Stemper, B. D., Yoganandan, N., Gennarelli, T. A., & Pintar, F. A. (2005). Localized cervical facet joint kinematics under physiological and whiplash loading. *Journal of Neurosurgery: Spine*, *3*(6), 471–476. https://doi.org/10.3171/spi.2005.3.6.0471

Stitt, D., Draper, N., Alexander, K., & Kabaliuk, N. (2021). Laboratory Validation of Instrumented Mouthguard for Use in Sport. *Sensors (Basel, Switzerland)*, *21*(18). https://doi.org/10.3390/S21186028

Stokes, K. A., Locke, D., Roberts, S., Henderson, L., Tucker, R., Ryan, D., & Kemp, S. (2021). Does reducing the height of the tackle through law change in elite men's rugby union (The Championship, England) reduce the incidence of concussion? A controlled study in 126 games. *British Journal of Sports Medicine*, *55*(4), 220–225. https://doi.org/10.1136/bjsports-2019-101557

Stokes, K., Kemp, S., Hudson, S., Anstiss, T., Brooks, J., Bryan, R., Cross, M., Jones, B., Henderson, L., Lee, M., Rossiter, M., West, S., McKay, C., & Williams, S. (2023). *Professional Rugby Injury Surveillance Project 2021-22*. https://keepyourbootson.co.uk/rugbysafe-toolkit/research/

Stokes, K., Roberts, S., Mckay, C., Kemp, S., & Faull-Brown, R. (2023). *Youth Rugby Injury Surveillance and Prevention Project 2021-22*. https://keepyourbootson.co.uk/rugbysafe-toolkit/research/

Su, J., Yu, X., Wang, X., Wang, Z., & Chao, G. (2024). Enhanced transfer learning with data augmentation. *Engineering Applications of Artificial Intelligence*, *129*. https://doi.org/10.1016/J.ENGAPPAI.2023.107602

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book.

Tabrizi, S. S., Pashazadeh, S., & Javani, V. (2020). Comparative Study of Table Tennis Forehand Strokes Classification Using Deep Learning and SVM. *IEEE Sensors Journal*, *20*(22), 13552–13561. https://doi.org/10.1109/JSEN.2020.3005443

Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*. *International Journal of Advanced Computer Science and Applications*, *11*(2). https://doi.org/10.14569/IJACSA.2020.0110277

Tooby, J., Till, K., Gardner, A., Stokes, K., Tierney, G., Weaving, D., Rowson, S., Ghajari, M., Emery, C., Bussey, M. D., & Jones, B. (2024). When to Pull the Trigger: Conceptual Considerations for Approximating Head Acceleration Events Using Instrumented Mouthguards. *Sports Medicine*, *54*(6), 1361–1369. https://doi.org/10.1007/s40279-024-02012-5

Tooby, J., Weaving, D., Al-Dawoud, M., & Tierney, G. (2022a). Quantification of Head Acceleration Events in Rugby League: An Instrumented Mouthguard and Video Analysis Pilot Study. *Sensors*, *22*(2), 584. https://doi.org/10.3390/s22020584

Tooby, J., Weaving, D., Al-Dawoud, M., & Tierney, G. (2022b). Quantification of Head Acceleration Events in Rugby League: An Instrumented Mouthguard and Video Analysis Pilot Study. *Sensors*, *22*(2), 584. https://doi.org/10.3390/s22020584

Tooby, J., Woodward, J., Tucker, R., Jones, B., Falvey, É., Salmon, D., Bussey, M. D., Starling, L., & Tierney, G. (2023). Instrumented Mouthguards in Elite-Level Men's and Women's Rugby Union: The Incidence and Propensity of Head Acceleration Events in Matches. *Sports Medicine*. https://doi.org/10.1007/s40279-023-01953-7

Tooby, J., Woodward, J., Tucker, R., Jones, B., Falvey, É., Salmon, D., Bussey, M. D., Starling, L., & Tierney, G. (2024). Instrumented Mouthguards in Elite-Level Men's and Women's Rugby Union: The Incidence and Propensity of Head Acceleration Events in Matches. *Sports Medicine*, *54*(5), 1327–1338. https://doi.org/10.1007/s40279-023-01953-7

Trotter, W. (1924). ON CERTAIN MINOR INJURIES OF THE BRAIN: Being the Annual Oration, Medical Society of London. *BMJ*, *1*(3306), 816–819. https://doi.org/10.1136/bmj.1.3306.816

Turner, R. C., Lucke-Wold, B. P., Robson, M. J., Lee, J. M., & Bailes, J. E. (2016). Alzheimer's disease and chronic traumatic encephalopathy: Distinct but possibly overlapping disease entities. *Brain Injury*, *30*(11), 1279–1292. https://doi.org/10.1080/02699052.2016.1193631

Um, T. T., Pfister, F. M. J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., & Kulic, D. (2017). Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, *2017-Janua*, 216–220. https://doi.org/10.1145/3136755.3136817

Ventresca, M., & Henne, K. (2020). *NFL concussion lawsuit payouts reveal how racial bias in science continues*. The Conversation. https://theconversation.com/nfl-concussion-lawsuit-payouts-reveal-how-racial-bias-in-science-continues-145987

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, *15*(12 December). https://doi.org/10.1371/JOURNAL.PONE.0243300

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walter, P. (2009). Accelerometer Limitations for Pyroshock Measurements. *Sound and Vibration*.

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*(December 2017), 3–11. https://doi.org/10.1016/j.patrec.2018.02.010

Wang, T., Kenny, R., & Wu, L. C. (2021). Head Impact Sensor Triggering Bias Introduced by Linear Acceleration Thresholding. *Annals of Biomedical Engineering*, *49*(12), 3189–3199. https://doi.org/10.1007/s10439-021-02868-y

Weir, B. S., & Hill, W. G. (2002). Estimating F-Statistics. *Annual Review of Genetics*, *36*(1), 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940

West, S. W., Starling, L., Kemp, S., Williams, S., Cross, M., Taylor, A., Brooks, J. H. M., & Stokes, K. A. (2021). Trends in match injury risk in professional male rugby union: a 16-season review of 10 851 match injuries in the English Premiership (2002–2019): the Professional Rugby Injury Surveillance Project. *British Journal of Sports Medicine*, *55*(12), 676–682. https://doi.org/10.1136/bjsports-2020-102529

Williams, E. M. P., Petrie, F. J., Pennington, T. N., Powell, D. R. L., Arora, H., Mackintosh, K. A., & Greybe, D. G. (2021). Sex differences in neck strength and head impact kinematics in university rugby union players. *European Journal of Sport Science*, 1–10. https://doi.org/10.1080/17461391.2021.1973573

Williams, S., Smith, K., Stokes, K., McKay, C., Kemp, S., Wojek, K., Jones, L., Bosshardt, C., Henderson, L., Byrne, A., Ross, E., Tyler, K., & Holmes, D. (2024). Women's Rugby Injury Surveillance Project (WRISP). *England Rugby*.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Woodhouse, L. N., Tallent, J., Patterson, S. D., & Waldron, M. (2022). International female rugby union players' anthropometric and physical performance characteristics: A five-year longitudinal analysis by individual positional groups.

Journal of Sports Sciences, *40*(4), 370–378. https://doi.org/10.1080/02640414.2021.1993656

World Rugby. (2019). *World Rugby Year in Review 2018*.

World Rugby. (2022, July 12). *World's top players to wear smart mouthguards at Rugby World Cup 2021 in landmark agreement to help reduce concussions*. Https://Www.World.Rugby/Keep-Rugby-Clean/News/731645/Worlds-Top-Players-to-Wear-Smart-Mouthguards-at-Rugby-World-Cup-2021-in-Landmark-Agreement-to-Help-Reduce-Concussions.

World Rugby. (2023, October 9). *World Rugby integrates smart mouthguard technology to the Head Injury Assessment as part of new phase of global player welfare measures*. Https://Www.World.Rugby/News/875212/World-Rugby-Integrates-Smart-Mouthguard-Technology-to-the-Head-Injury-Assessment-as-Part-of-New-Phase-of-Global-Player-Welfare-Measures.

Wright, F., Docherty, P. D., Williams, E., Greybe, D., Arora, H., & Kabaliuk, N. (2021). An in-silico study of the effect of non-linear skin dynamics on skin-mounted accelerometer inference of skull motion. *Biomedical Signal Processing and Control*, *70*, 102986. https://doi.org/10.1016/j.bspc.2021.102986

Wu, L. C., Kuo, C., Loza, J., Kurt, M., Laksari, K., Yanez, L. Z., Senif, D., Anderson, S. C., Miller, L. E., Urban, J. E., Stitzel, J. D., & Camarillo, D. B. (2018). Detection of American Football Head Impacts Using Biomechanical Features and Support Vector Machine Classification. *Scientific Reports*, *8*(1), 1–14. https://doi.org/10.1038/s41598-017-17864-3

Wu, L. C., Nangia, V., Bui, K., Hammoor, B., Kurt, M., Hernandez, F., Kuo, C., & Camarillo, D. B. (2016). In Vivo Evaluation of Wearable Head Impact Sensors. *Annals of Biomedical Engineering*, *44*(4), 1234–1245. https://doi.org/10.1007/s10439-015-1423-3

Wu, L. C., Zarnescu, L., Nangia, V., Cam, B., & Camarillo, D. B. (2014). A head impact detection system using SVM classification and proximity sensing in an instrumented mouthguard. *IEEE Transactions on Biomedical Engineering*, *61*(11), 2659–2668. https://doi.org/10.1109/TBME.2014.2320153

Xiong, Y., GU, Q., PETERSON, P. L., MUIZELAAR, J. P., & LEE, C. P. (1997). Mitochondrial Dysfunction and Calcium Perturbation Induced by Traumatic Brain Injury. *Journal of Neurotrauma*, *14*(1), 23–34. https://doi.org/10.1089/neu.1997.14.23

Yamada, M. (2006). Wavelets: Applications. In *Encyclopedia of Mathematical Physics* (pp. 420–426). Elsevier. https://doi.org/10.1016/B0-12-512666-2/00242-X

Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., & Maybank, S. J. (2019). Asymmetric 3D Convolutional Neural Networks for action recognition. *Pattern Recognition*, *85*, 1–12. https://doi.org/10.1016/j.patcog.2018.07.028

Yan-yan, S. O. N. G., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

Yoganandan, N., Pintar, F. A., Zhang, J., & Baisden, J. L. (2009). Physical properties of the human head: Mass, center of gravity and moment of inertia. *Journal of Biomechanics*, *42*(9), 1177–1192. https://doi.org/10.1016/j.jbiomech.2009.03.029

Zhang, D. (2019). *Wavelet Transform* (pp. 35–44). https://doi.org/10.1007/978-3-030-17989-2_3

Zhang, L., Yang, K. H., & King, A. I. (2004). A Proposed Injury Threshold for Mild Traumatic Brain Injury. *Journal of Biomechanical Engineering*, *126*(2), 226–236. https://doi.org/10.1115/1.1691446

Zhang, W., Su, W., Hu, Z., Lu, J., Yu, Y., & Chow, K. (2022, February 10). *A Practical Guide to Support Vector Machines (SVM)*. Medium. https://medium.com/sfu-cspmp/a-practical-guide-to-support-vector-machines-svm-ccd6a4d4dd04

Zhao, Z., Anand, R., & Wang, M. (2019a). Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. *Proceedings - 2019 IEEE International Conference on Data Science and*

Advanced Analytics, DSAA 2019, 442–452. https://doi.org/10.1109/DSAA.2019.00059

Zhao, Z., Anand, R., & Wang, M. (2019b). Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. *International Conference on Data Science and Advanced Analytics*, 442–452. https://doi.org/10.1109/DSAA.2019.00059

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (1st ed.). O'Reilly Media, Inc.

12 **Appendices**

Appendix A: The data after conversion to features for the DHC vs NDC classification challenge.

Appendix B: The cell block containing the required features for the DHC vs NDC classification challenge.

```
# Import necessary libraries
import numpy as np
import os
import glob
import pywt
import pandas as pd
from scipy import signal
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.decomposition import PCA

# Define the frequency range for the PSD and WT feature extraction
freq_range = np.arange(10, 260, 10)

# Define the variables/columns for different feature types
vecs = ["XL", "YL", "ZL", "RL", "XV", "YV", "ZV", "RV", "XA", "YA", "ZA", "RA"]
res = ["RL", "RV", "RA"]

# Define the features dictionary that maps feature types to specific variables
features = {
    "PSD": {
        "XL": freq_range, "YL": freq_range, "ZL": freq_range,
        "RL": freq_range, "XV": freq_range, "YV": freq_range,
        "ZV": freq_range, "RV": freq_range,
        "XA": freq_range, "YA": freq_range, "ZA": freq_range, "RA": freq_range
    },
    "WT": {
        "XL": freq_range, "YL": freq_range, "ZL": freq_range,
        "RL": freq_range, "XV": freq_range, "YV": freq_range,
        "ZV": freq_range, "RV": freq_range,
        "XA": freq_range, "YA": freq_range, "ZA": freq_range, "RA": freq_range
    },
    "Pulse Width": vecs,  # Pulse Width for the listed vector variables
    "Pulse Prom": vecs,   # Pulse Prominence for the listed vector variables
    "Pulse Num": vecs,    # Pulse Count for the listed vector variables
    "Jerk": vecs,         # Jerk (rate of change of acceleration) for the listed vector variables
    "Snap": vecs,         # Snap (rate of change of jerk) for the listed vector variables
    "WT_T0": {
        "XL": freq_range, "YL": freq_range, "ZL": freq_range,
        "RL": freq_range, "XV": freq_range, "YV": freq_range,
        "ZV": freq_range, "RV": freq_range,
        "XA": freq_range, "YA": freq_range, "ZA": freq_range, "RA": freq_range
    },
    "PCA": vecs,          # Principal Component Analysis for the listed vector variables
    "Mean": vecs,         # Mean for the listed vector variables
    "Variance": vecs,     # Variance for the listed vector variables
    "Std_dev": vecs,      # Standard Deviation for the listed vector variables
    "Max": vecs,          # Maximum value for the listed vector variables
    "AUC": res            # Area Under Curve for the listed result variables (RL, RV, RA)
}

```

Appendix C: Part 1 of the code used to extract features for the DHC vs NDC classification challenge.

```python
import glob
import os
import numpy as np
import pandas as pd
from scipy import signal
import pywt
from sklearn.decomposition import PCA

def Feature_Extraction(features, label, path_in):
    """
    This function extracts a set of features from CSV files in the specified directory,
    processes the data according to the feature types provided,
    and returns a DataFrame containing the extracted features along with a label.
    """

    # Initialize variables to hold extracted data and file details
    count = 0
    output_array = []  # Stores extracted feature values
    dataframes = []  # Stores data from CSV files
    f_names = []  # List to store feature names

    # Generate a list of all feature names based on the input `features` dictionary
    for feature_key in features.keys():
        if isinstance(features[feature_key], list):
            for item in features[feature_key]:
                f_names.append(f"{feature_key}_{item}")
        else:
            for sub_key in features[feature_key].keys():
                for sub_item in features[feature_key][sub_key]:
                    f_names.append(f"{feature_key}_{sub_key}_{sub_item}")

    # List all CSV files in the specified input directory
    file_list = glob.glob(os.path.join(path_in, "*.csv"))

    # Constants for signal processing
    fs = 3200  # Sampling frequency
    N = 160  # Signal length (example)
    nperseg = 160  # Segment length for FFT
    sampling_period = 0.0003125  # Sampling period
    f_corr_coef = 1  # Correction coefficient for feature scaling

    # Indexes of columns in the CSV to be used
    index = np.arange(1, 18, 1)

    # Loop through each file and process the data
    for file_path in file_list:
        # Read the CSV file and split the data into three segments
        name = os.path.basename(file_path)
        df = pd.read_csv(file_path, usecols=index)

        dfAA, dfAV, dfLA = np.array(df.iloc[:, 0:4]), np.array(df.iloc[:, 4:8]), np.array(df.iloc[:, 8:12])
        df = np.concatenate((dfLA, dfAV, dfAA), axis=1)

        # Add the filename to the dataframe and append it to the list
        df = pd.DataFrame(df)
        df['File Name'] = name
        dataframes.append(df)

    # Feature extraction loop: For each dataframe, calculate features
    for data_df in dataframes:
        if not data_df.isnull().values.any():
            # Rename columns for consistency
            cols = ["XL", "YL", "ZL", "RL", "XV", "YV", "ZV", "RV", "XA", "YA", "ZA", "RA", "Col_N"]
            data_df.columns = cols

            # Extract the feature name
            fn = data_df['Col_N'][0]
            f_vals = []

            # Loop through each feature type and calculate the corresponding feature values
            for f_type in features.keys():
                if f_type == "Pulse Width":
                    for feature in features["Pulse Width"]:
                        pulse_w = []
                        ABS = abs(data_df[feature])
                        for val in ABS:
                            if val > np.mean(ABS):
                                indices = list(ABS).index(val)
                                pulse_w.append(indices)

                        previous = 0
                        max_list = 1
                        top_len = 1
                        for j in pulse_w:
                            if j == previous + 1:
                                previous = j
                                max_list += 1
                                top_len = max(top_len, max_list)
                            else:
                                previous = j
                                max_list = 1
                        f_vals.append(top_len * f_corr_coef)
```

Appendix D: Part 2 of the code used to extract features for the DHC vs NDC classification challenge.

```
                elif f_type == "Pulse Prom":
                    for feature in features["Pulse Prom"]:
                        no, lib = signal.find_peaks(data_df[feature], prominence=(0))
                        if len(lib["prominences"]) == 0:
                            f_vals.append(0)
                        else:
                            pk_prom = max(lib["prominences"])
                            f_vals.append(pk_prom)

                elif f_type == "Pulse Num":
                    for feature in features["Pulse Num"]:
                        no, lib = signal.find_peaks(data_df[feature])
                        num_pk = len(no)
                        f_vals.append(num_pk)

                elif f_type == "Jerk":
                    for feature in features["Jerk"]:
                        f_vals.append(max(np.diff(data_df[feature], n=1)))

                elif f_type == "Snap":
                    for feature in features["Snap"]:
                        f_vals.append(max(np.diff(data_df[feature], n=2)))

                elif f_type == "PSD":
                    for feature in features["PSD"]:
                        f, Pxx_spec = signal.welch(data_df[feature], fs, window='flattop', nperseg=104)
                        for j in features["PSD"][feature]:
                            f_vals.append(Pxx_spec[round(int(j) / 9.19230769)])

                elif f_type == "WT":
                    for feature in features["WT"]:
                        wt_fqs = [81.25 / (j * 0.1) for j in features["WT"][feature]]
                        impact_max = feature.index(max(feature))
                        coef, freqs = pywt.cwt(data_df[feature], wt_fqs, 'mexh', sampling_period=sampling_period)
                        f_vals.extend(list(coef[:, impact_max]))

                elif f_type == "WT_T0":
                    for feature in features["WT_T0"]:
                        wt_fqs = [81.25 / (j * 0.1) for j in features["WT_T0"][feature]]
                        impact_0 = 0
                        coef, freqs = pywt.cwt(data_df[feature], wt_fqs, 'mexh', sampling_period=sampling_period)
                        f_vals.extend(list(coef[:, impact_0]))

                elif f_type == "PCA":
                    pca_img = PCA(n_components=1)
                    for feature in features["PCA"]:
                        PC_img = pca_img.fit_transform(np.array(data_df[feature]).reshape(-1, 1))
                        f_vals.append(PC_img[0][0])

                elif f_type == "Mean":
                    for feature in features["Mean"]:
                        f_vals.append(np.mean(data_df[feature]))

                elif f_type == "Variance":
                    for feature in features["Variance"]:
                        f_vals.append(np.var(data_df[feature]))

                elif f_type == "Std_dev":
                    for feature in features["Std_dev"]:
                        f_vals.append(np.std(data_df[feature]))

                elif f_type == "Max":
                    for feature in features["Max"]:
                        f_vals.append(max(abs(data_df[feature])))

                elif f_type == "AUC":
                    for feature in features["AUC"]:
                        f_vals.append(np.trapz(data_df[feature]))

            # Append the calculated feature values to the output array
            f_vals_array = np.array(f_vals, dtype=object).reshape(1, -1)
            if count == 0:
                output_array = f_vals_array
            else:
                output_array = np.concatenate((output_array, f_vals_array), axis=0)
            count += 1

    # Create the final labeled output
    m, n = output_array.shape
    labels = np.ones((m, 1)) * label
    DA_Labelled = np.concatenate([output_array, labels], axis=1)

    # Add the 'Label' column to the feature names
    f_names.append("Label")

    # Create the DataFrame with labeled features
    output_df = pd.DataFrame(data=DA_Labelled, columns=f_names)

    return output_df

```

Appendix E: Part 1 of the code used to train classifiers and report performance for the DHC vs NDC classification challenge.

```python
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFE
from sklearn.ensemble import IsolationForest
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.metrics import recall_score, roc_auc_score, confusion_matrix
from mrmr import mrmr_classif

def Train_Test(clf, its, train_csv, test_csv):
    """
    Function to train and test a classifier with feature selection,
    outlier removal, oversampling, and evaluation metrics.

    Parameters:
    clf (object): Classifier model (e.g., RandomForest, SVM).
    its (int): Number of features to consider.
    train_csv (str): Path to the training CSV file.
    test_csv (str): Path to the testing CSV file.

    Returns:
    selected_features (list): List of selected features after feature selection.
    train_scores_AR (list): List of recall scores for training data.
    test_scores_AR (list): List of recall scores for testing data.
    train_scores_AUROC (list): List of AUROC scores for training data.
    test_scores_AUROC (list): List of AUROC scores for testing data.
    """

    # Initialize lists to store scores for evaluation metrics
    val_scores_XGB = []
    train_scores_AR = []
    test_scores_AR = []
    train_scores_AUROC = []
    test_scores_AUROC = []

    # Create a list of feature counts to iterate over (even numbers from 2 to 'its' + 2)
    ks = np.arange(2, its + 2, 2)

    # Load the training and test datasets
    train = pd.read_csv(train_csv)
    test = pd.read_csv(test_csv)

    # Split the datasets into features (X) and labels (y)
    X_train, y_train = train.iloc[:, :-1], train.iloc[:, -1]
    X_test, y_test = test.iloc[:, :-1], test.iloc[:, -1]

    # Feature selection using MRMR (Minimum Redundancy Maximum Relevance)
    selected_features = mrmr_classif(X=X_train, y=y_train, K=int(its))
    X_train_mrmr = X_train[selected_features]

    print("Selected Features:", selected_features)

    # Iterate over the number of features to keep (k) and evaluate models
```

Appendix F: Part 2 of the code used to train classifiers and report performance for the DHC vs NDC classification challenge.

```
    # Iterate over the number of features to keep (k) and evaluate models
    for k in ks:
        # Select the top 'k' features for training
        X_train_mrmr_k = X_train_mrmr.iloc[:, :k]
        y_train_k = y_train

        # Prepare test data (using all features)
        X_test_k = X_test
        y_test_k = y_test

        # Eliminate features using Recursive Feature Elimination (RFE)
        rfe = RFE(clf)
        rfe.fit(X_train_mrmr_k, y_train_k)

        # Select the features that are kept by RFE
        X_train_RFE = X_train_mrmr_k.loc[:, rfe.support_]
        X_test_RFE = X_test_k[X_train_RFE.columns]

        # Eliminate outliers using Isolation Forest
        isf = IsolationForest(n_jobs=-1, random_state=1)
        isf.fit(X_train_RFE, y_train)

        # Identify outliers (-1 are outliers)
        remove = [row for row, val in pd.DataFrame(isf.predict(X_train_RFE)).iterrows() if val[0] == -1]

        # Remove outliers from the training data
        X_train_EDP = X_train_RFE.drop(remove)
        y_train_EDP = y_train_k.drop(remove)

        # Oversample the training data to handle class imbalance using SMOTE
        oversample = SMOTE()
        X_train_OvS, y_train_OvS = oversample.fit_resample(X_train_EDP, y_train_EDP)

        # Standardize the data using StandardScaler
        scaler = StandardScaler().fit(X_train_OvS)
        X_train_Scaled = pd.DataFrame(data=scaler.transform(X_train_OvS), columns=X_train_OvS.columns)
        X_test_Scaled = pd.DataFrame(data=scaler.transform(X_test_RFE), columns=X_train_OvS.columns)

        # Train the classifier on the scaled training data
        clf.fit(X_train_Scaled, y_train_OvS)

        # Predict probabilities and labels for training and test data
        train_proba = clf.predict_proba(X_train_Scaled)[:, 1]
        train_pred = clf.predict(X_train_Scaled)
        test_proba = clf.predict_proba(X_test_Scaled)[:, 1]
        test_pred = clf.predict(X_test_Scaled)

        # Evaluate the model performance using recall score and AUROC
        train_AR = recall_score(train_pred, y_train_OvS, average='macro')
        test_AR = recall_score(test_pred, y_test_k, average='macro')
        train_AUROC = roc_auc_score(y_train_OvS, train_proba)
        test_AUROC = roc_auc_score(y_test_k, test_proba)

        # Store the evaluation scores for later analysis
        train_scores_AR.append(train_AR)
        test_scores_AR.append(test_AR)
        train_scores_AUROC.append(train_AUROC)
        test_scores_AUROC.append(test_AUROC)

        # Confusion matrix for training and testing data (normalized)
        results_train = confusion_matrix(y_train_OvS, train_pred)
        results_train_normalised = results_train / results_train.astype(np.float).sum(axis=1)[:, np.newaxis]

        results_test = confusion_matrix(y_test_k, test_pred)
        results_test_normalised = results_test / results_test.astype(np.float).sum(axis=1)[:, np.newaxis]

        # Plot the confusion matrices using heatmaps
        fig, axes = plt.subplots(1, 2)

        sns.heatmap(results_train_normalised, annot=True, cmap='Blues', ax=axes[0])
        sns.heatmap(results_test_normalised, annot=True, cmap='Blues', ax=axes[1])

        plt.show()

    # Return the selected features and evaluation metrics for later analysis
    return selected_features, train_scores_AR, test_scores_AR, train_scores_AUROC, test_scores_AUROC

```