MACHINE LEARNING APPLICATION IN PREDICTING ANTERIOR CRUCIATE LIGAMENT INJURY AMONG BASKETBALL PLAYERS

GUO LONGFEI

UNIVERSITI SAINS MALAYSIA

2025

MACHINE LEARNING APPLICATION IN PREDICTING ANTERIOR CRUCIATE LIGAMENT INJURY AMONG BASKETBALL PLAYERS

by

GUO LONGFEI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

January 2025

ACKNOWLEDGEMENT

Firstly, I express my profound gratitude to my supervisor, Assoc Prof. Ts Dr. Shazlin Shaharudin, and my co-supervisors, Assoc Prof. Dr Loh Wei Ping and Assoc Prof Dr. Fatanah Mohamad Suhaimi. Despite the challenges posed by the pandemic, which initially limited our meetings to online interactions, Associate Professor Ts.Dr. Shazlin helped me navigate the direction of my thesis. When my work hit a roadblock, she provided invaluable guidance. Under my supervisors' tutelage, I mastered more research skills including writing, literature reviews, and selecting appropriate journals for publication. With their help, I have reached this point in my journey. I also wish to express my gratitude to other lecturers and students in USM, who deepened my understanding of Malaysia as a multicultural nation. Their enthusiasm and friendships, especially within the School of Health Sciences, have been invaluable. Additionally, I must express my gratitude to Professor Zhi Lei Cui, Professor Rui Cao, Professor Xin Wen from Taiyuan University of Technology, and Professor Zhong Yuan Ding from Xinzhou Teachers University. Their significant support with experimental equipment, participant recruitment, and machine learning modelling has been invaluable. I am also thankful to He-Rui Jie and Wang Zhaohui for their insights on machine learning methodologies—Special thanks to Zhizhi Van for his diligent efforts in implementing my experiments and data collection. I greatly thank my parents, my wife, and my children, whose understanding and support have been crucial. Without them, I could not have completed my studies. Lastly, I want to express my sincere gratitude to everyone who has supported me throughout my doctoral studies.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ACL	Anterior Cruciate Ligament
PCL	Posterior Cruciate Ligament
RF	Random Forest
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting
LR	Logistic Regression
LESS	Landing Error Scoring System
CMAS	Cutting Movement Assessment Score
BMI	Body Mass Index
ML	Machine Learning
SMOTE	Synthetic Minority Over-Sampling Technique
SHAP	SHapley Additive exPlanations
VIF	Variance Inflation Factor
MRI	Magnetic Resonance Imaging
1RM	One-Repetition Max
mRFD	Maximum Rate of Force Development
COM	Centre of Mass
COP	Centre of Pressure
Q Angle	Quadriceps Angles
MTD	Medial Tibial Depth
LPTS	Lateral Posterior Tibial Slope

TS	Tibial Slope
LFCR	Lateral Femoral Condyle Ratio
ST	Semitendinosus
VL	Vastus Lateralis
PS	Peak Stance
FMS	Functional Movement Screen
SEBT	Star Excursion Balance Test
YBT	Y Balance Test
ICC	Intra-class Correlation Coefficients
RDL	Romanian Dead Lift
DLH	Dominant Leg Hop
SJ	Squat Jump
CMJ	Counter Movement Jump
DJ	Drop Jump
COD	Change of Direction
DVJ	Drop vertical jump
SLDL	Single-leg Drop Landing
USSC	Unanticipated Side-step Cutting
CNN	Convolutional Neural Network
GMM	Gaussian Mixture Model
GPC	Gaussian Process Classification
TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives

TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CV	Cross-validation
K-score	Cohen's Kappa
GRF	Ground Reaction Force
ES	Emergency Stop
IA	Initial Acceleration
SC	Side-Cutting
IQR	Interquartile Range
EMG	Electromyography
MVC	Maximum Voluntary Contraction
CSCS	Certified Strength and Conditioning Specialist
RSS	Residual Sum of Squares
ACWR	Acute Chronic Workload Ratio

LIST OF APPENDICES

APLIKASI PEMBELAJARAN MESIN UNTUK MERAMAL KECEDERAAN LIGAMEN ANTERIOR CRUCIATE DALAM KALANGAN PEMAIN BOLA KERANJANG

ABSTRAK

Kecederaan Ligamen Anterior Cruciate (ACL) adalah antara kecederaan yang paling kerap berlaku dalam kalangan atlet, yang memberi kesan besar kepada prestasi kompetitif mereka. Mencegah kecederaan ACL adalah mencabar kerana sifatnya yang pelibagai faktor. Teknik perlombongan data berasaskan pembelajaran mesin telah menunjukkan potensi besar dalam mengenal pasti faktor risiko yang berkaitan dengan kecederaan ACL. Kajian ini bertujuan untuk menilai keupayaan meramal faktor-faktor ini menggunakan model pembelajaran mesin. Data mengenai profil atlet, fungsi fizikal, kualiti khusus, analisis pergerakan tiga dimensi, dan elektromiografi serentak telah dikumpulkan secara prospektif daripada 104 pemain bola keranjang lelaki. Susulan selama satu tahun dijalankan untuk memantau kecederaan ACL, dengan 11 pemain dikenal pasti mengalami kecederaan. Empat algoritma pembelajaran mesin — Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), dan Logistic Regression (LR) — dibangunkan untuk meramalkan kecederaan ACL. Model terbaik dipilih berdasarkan purata kawasan di bawah lengkung ciri operasi penerima (AUC-ROC) daripada 10 ulangan validasi silang dan digunakan bersama Shapley Additive exPlanations (SHAP) untuk menganalisis faktor risiko. Keputusan menunjukkan nilai AUC-ROC sedikit berbeza antara ulangan dan kaedah (0.66-0.80), dengan pengelas terbaik

adalah RF. Analisis SHAP mengenal pasti ciri-ciri utama dengan nilai ramalan tertinggi untuk kecederaan ACL semasa pergerakan sukan tertentu. Fasa Berhenti Kecemasan: Peningkatan momen fleksi lutut, daya tindak balas tanah posterior, sudut fleksi lutut, dan pengaktifan berlebihan kuadrisep lateral serta otot rectus femoris. Fasa Pecutan Awal: Peningkatan tork putaran dalaman lutut dan tekanan lateral pada anggota kaki. Fasa Pemotongan Sisi: Penurunan kecondongan tibial dan sudut fleksi pinggul, peningkatan sudut inversi pergelangan kaki, momen eversi pergelangan kaki, dan pengaktifan berlebihan otot paha lateral. Selain itu, kestabilan yang lemah pada kaki bukan dominan, prestasi Squat Jump yang rendah, beban latihan melebihi 15 jam seminggu, dan sejarah kecederaan sebelum ini adalah peramal ketara untuk kecederaan ACL. Kajian ini menekankan keberkesanan model Pembelajaran Mesin dalam meramalkan kecederaan ACL, dengan mengutamakan metrik biomekanik, atribut fungsi, dan faktor sejarah sebagai peramal penting untuk strategi pencegahan yang disasarkan.

MACHINE LEARNING APPLICATION IN PREDICTING ANTERIOR CRUCIATE LIGAMENT INJURY AMONG BASKETBALL PLAYERS

ABSTRACT

Anterior cruciate ligament (ACL) injury is among the most prevalent injuries in athletes, significantly impacting their competitive performance. Preventing ACL injury is challenging due to their multifactorial nature. Machine learning-based data mining techniques have shown significant potential in identifying risk factors associated with ACL injury. This study aimed to assess the predictive capability of these features using machine learning models. Data on athlete's profile, physical function, specialized qualities, three-dimensional movement analysis, and simultaneous electromyography were prospectively collected from 104 male basketball players. A one-year follow-up was conducted to monitor ACL injury, identifying n=11 injured players. Four machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Logistic Regression (LR)—were developed to predict ACL injury. The optimal model was selected based on the mean area under the receiver operating characteristic curve (AUC-ROC) across 10 cross-validation runs and was used with Shapley Additive exPlanations to analyze the risk factors. The results show that AUC-ROC values varied slightly across repetitions and methods (0.66-0.80), the best classifier was RF. SHAP analysis identified key feature with the highest predictive value for ACL injury during specific sports motions. Emergency Stop phase:

Increased knee flexion moment, posterior ground reaction forces, knee flexion angle, and overactivation of the lateral quadriceps and rectus femoris. Initial Acceleration phase: Elevated knee internal rotation torque and lateral stress on the lower limbs. Side-Cutting phase: Reduced tibial inclination and hip flexion angles, increased ankle inversion angle, ankle eversion moment, and excessive lateral thigh muscle activation. Furthermore, poor stability in the non-dominant leg, weak Squat Jump performance, training loads exceeding 15 hours per week, and prior injury history were significant ACL injury predictors. This study emphasizes the Machine Learning model's effectiveness in predicting ACL injury, highlighting biomechanical metrics, functional attributes, and historical feature as critical predictors for targeted prevention strategies.

CHAPTER 1

INTRODUCTION

1.1 Background of the study

Basketball is a high-intensity competitive sport, which requires athletes to dribble, breakthrough and shooting while running at a high speed. Due to the game intensity and prolonged training, basketball players are prone to sports injuries. Among them, Anterior Cruciate Ligament (ACL) injury is one of the most common sports injuries of basketball players (Escamilla et al., 2012; Westermann et al., 2019). According to an epidemiological study, the incidence rate of ACL injury in team sports in the United States is 0.15% to 0.36%, annually (Rosa et al., 2014). Also in China, ACL-related injuries in basketball players accounted for approximately 22% of the total injuries, and the number of ACL-related injuries is increasing every year (Lu & Zhang., 2014).

ACL is a complex structure that can withstand multiaxial stress and variable tensile strain to prevent leg over-extension and knee valgus. The high-risk actions of ACL injury include side cutting, jumping, and sudden deceleration, which occur repeatedly in basketball training and competition (Micheo et al., 2010). ACL injury will seriously affect the competitive state of athletes. Even with ACL reconstruction, it will take 4-4.6 seasons to recover (Mihata et al., 2006). Moreover, the surgery and subsequent rehabilitation are time-consuming and expensive, which will be a heavy burden on the patients' physiology, psychology and finance. As the core talent pool

of basketball, strategies for ACL injury prevention among the collegiate basketball teams should be considered.

"Prevention is greater than treatment" has become a consensus in dealing with non-contact ACL injury (Zhang et al., 2019). Currently, a variety of tests are available to assess athletes for features associated with their susceptibility and risk of ACL injury, such as Landing Error Scoring System (LESS) (Beutler et al., 2009), and Cutting Movement Assessment Score (CMAS） (Hughes & Watkins, 2006). Although, standardized movement tests are highly applicable and simple to execute, but most studies evaluated a single biomechanical factor of ACL risk injury for example dynamic knee valgus (Dai et al., 2012) or landing technique (Dai et al., 2015). Despite that, other potential features of ACL injury also include lower extremity or core muscle strength deficits (Raschner et al., 2012), lack of balance and joint laxity (Oshima et al., 2018), and increased body mass index (BMI) (Cronström et al., 2023). Therefore, predicting ACL injury requires a comprehensive approach that integrates multidimensional movement patterns and dynamic system interactions to account for the complex interplay of biomechanical, muscular, and physiological risk factors.

Machine learning (ML) technology, represented by simulating human learning behaviour, has become more mature, and widely used in various fields of Artificial Intelligence. Sports injuries and rehabilitation have recently benefited from ML applications, as ML models can capture the interaction of multiple predictors. For example, Taborri et al. (2021) identified the risk of ACL injury in 39 basketball players by machine learning algorithm. Their results showed that the Support Vector Machine (SVM) algorithm could achieve 96% accuracy, demonstrating a high

predictive effect (Taborri et al., 2021). However, this study is limited by a small sample size, a narrow selection of influencing features, and a lack of robust validation for the model's results, which may impact the generalizability and reliability of its findings.

Machine learning is a powerful tool for automated decision-making, but its effectiveness in injury prediction modeling depends on several critical features. First and foremost is the quality of data, as the reliability and performance of ML models are directly influenced by the accuracy, relevance, and comprehensiveness of the input data (Jain et al., 2020). This current study used prospective experimental data, ensured dataset reliability through Lachman clinical tests, and performed differential testing to identify risk factors linked to ACL injury. Secondly, addressing the challenge of data imbalance is essential. In sports injury prediction using ML, the injury samples are usually a very small fraction of the total number. Therefore, the class-imbalance data can lead to overfitting (Jauhiainen et al., 2021). The current study employs a combination of Gaussian noise and Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalance in the data. This approach not only simulates the noise of real data but also ensures that the model has sufficient samples to learn the complex patterns in the data. Moreover, the use of cross-validation, the most commonly used way to estimate model generalization ability in many fields, introduces randomness to the analysis can be effective in preventing data overfitting (Forman & Scholz, 2010). While the primary goal of machine learning is often to achieve high predictive accuracy on independent test data, this study also emphasized the importance of ensuring that the models are interpretable and their predictions are explainable. Hence, to interpret and visualize the output of each model, we used the SHapley Additive exPlanations (SHAP) approach

(Lundberg et al., 2018) because it provides a better understanding of the impact of different features on the model results. Finally, to ensure robust model performance, this study employed multiple randomized repetitions of experiments to enhance result stability and credibility while minimizing the influence of chance on predictions. A comprehensive evaluation of four machine learning algorithms—Random Forest, Logistic Regression, SVM, and XGBoost—was conducted using various performance metrics such as accuracy, precision, recall, F1 score, and AUC. These algorithms were selected for their specific strengths: Random Forest and XGBoost excel in handling complex nonlinear relationships, Logistic Regression is valued for its simplicity and interpretability, and SVM performs well in high-dimensional spaces. This approach allowed for the identification of the most suitable model tailored to the research task.

Based on the above, this study focuses on male basketball players and incorporates athlete's profile, physical function features, physical fitness tests, and biomechanical and electromyographic data collected during unanticipated lateral cutting maneuvers. Then, a multidimensional machine learning model for predicting ACL injury was developed. This study addresses the limitations of existing research in data integration, predictive performance, and model interpretability. It provides scientific evidence and theoretical support for the early identification and effective prevention of ACL injury.

1.2 Theoretical framework

The theoretical framework provides comprehensive understanding for ACL injury risk identification, data analysis, and model development through the integration and application of multidisciplinary theories. The framework not only

facilitates the achievement of research objectives but also lays a foundation for the scientific advancement of injury prediction and prevention. The specific theoretical components are as follows:

1.2.1 Multifactorial Model of Injury

The multifactorial model of sports injury originates from the four-step framework for injury prevention proposed by Van Mechelen et al. (1992). This framework systematically addresses injury identification, causal exploration, prevention strategy development, and outcome evaluation from an epidemiological perspective, laying a foundation for subsequent research. Meeuwisse (1994) further advanced this field by proposing the multifactorial model of sports injury, emphasizing that the interplay between intrinsic features and extrinsic features jointly determines injury risk. In 2007, the dynamic recursive model was proposed, which highlights the accumulation and dynamic variation of risk factors combined with the final triggering event may collectively drive injury occurrence (Meeuwisse et al., 2007). Bahr and Krosshaug (2005) introduced the concept of the injury causation chain, systematically breaking down the entire injury process for the first time—from individual baseline risk to acute precipitating features and triggering events. Using biomechanical techniques, they revealed key mechanisms underlying injuries, providing a basis for effective prevention. Entering the era of complex systems theory, Bittencourt et al. (2016) adopted a complex systems model perspective, viewing sports injuries as the result of nonlinear dynamic interactions among multidimensional features. Through network analysis, they uncovered the intricate relationships between these features, driving the transition from traditional causal analysis to precise prediction.

Based on the developmental trajectory of the multifactorial model of sports injuries, ACL injury research should adopt a multidimensional perspective to comprehensively uncover its complex mechanisms. Core risk factors include intrinsic features such as anatomical characteristics, physiological traits, and neuromuscular control capacity, which often determine an individual's baseline susceptibility to injury (Shultz et al., 2012). Secondly, extrinsic features related to environmental or external conditions significantly influence the occurrence of injuries. These include training and competition intensity, frequency, playing surfaces, and equipment (Alentorn-Geli et al., 2009). Thirdly, trigger events. Biomechanical imbalances during high-risk actions or events are often the direct causes of injury, such as insufficient knee flexion angle during jump landings and excessive knee shear forces during high-speed lateral movements. Consequently, biomechanical metrics have become the most commonly used predictors of ACL injury (Yu & Garrett, 2007). Furthermore, integrating complex systems theory is necessary to analyze the interactions among multidimensional features.

1.2.2. Theoretical Framework for Machine Learning in Injury Prediction

The core of machine learning in injury prediction lies in multidimensional data classification, where input features are analyzed to determine whether an athlete is at high risk. In the 1950s, the concept of using machine learning for prediction was proposed, laying the philosophical foundation for the field (Turing, 1950). Then, rule-based chess program demonstrated the potential of decision-making through pattern recognition, marking the prototype of supervised learning (Samuel, 1959). Next, statistical learning theory and support vector machines (SVM) provided the mathematical foundation for small-sample learning and high-dimensional modeling,

forming an early framework for injury risk classification tasks (Vapni, 2000). However, these models have limitations in handling nonlinear dynamic interactions. As machine learning shifted its focus to data-driven model optimization, Random Forest was proposed to enhance model robustness and generalization through ensemble learning methods (Breiman, 2001). Meanwhile, the decision tree model was developed to decomposed complex problems into multiple simple decision paths, significantly improving the model's interpretability and applicability (Quinlan, 2014). During this phase, machine learning was gradually applied to sports injury prediction, focusing on identifying specific high-risk movements, such as assessing knee joint mechanics during dynamic activities.

The interpretability of machine learning models has also become a key research focus. The LIME method enhances model transparency through local explanations, particularly in multivariable interaction analysis (Das et al., 2019). Meanwhile, SHAP tool was developed based on game theory, quantifying the contribution of each feature to the model’s predictions and enhancing prediction transparency. In the field of sports injury prediction, these tools help to interpret the specific impact of key features on injury risk (Lundberg & Lee, 2017).

Overall, the development of machine learning prediction theories has evolved from early probabilistic models to advanced models capable of capturing nonlinear dynamic relationships and multifactorial interactions, providing a theoretical foundation and practical support for precise sports injury prediction.

1.3 Conceptual Framework

This study's conceptual framework systematically illustrates the entire process, from identifying risk factors for ACL injury and data collection to building predictive models.

1.3.1 Classification of Risk factors for ACL Injury

Risk factors for ACL injury include anatomical structure, hormonal levels, and environmental features, which are typically difficult to modify and are defined as invariable features. In contrast, biomechanical and neuromuscular features related to ACL injury—such as biomechanics, joint flexibility, core stability, body composition, and sport-specific qualities—can be improved through training and are classified as variable features, which are the focus of this study.

1.3.2 Risk Data Collection and Injury Quantification for ACL Injury

Risk factors were identified through a literature review. Then, biomechanical data were collected during unanticipated stop-and-cut basketball maneuvers. Next, joint flexibility and stability were measured using the FMS test while the core stability was assessed through the Y-Balance test. The sport-specific qualities were quantified using strength, explosiveness, and agility tests. Finally, the athlete’s Profile was gathered through questionnaires. A 12-month follow-up was conducted post-testing, with ACL injury diagnosed using the Lachman test and MRI.

1.3.3 Construction of the ACL Injury Prediction Model

(1) Feature Selection: Independent sample t-tests were performed to include features that showed significant differences ($P < 0.05$) between injured and non-injured athletes. Feature selection reduces model complexity, enhances generalization, and ensures the relevance of the input features.

(2) VIF: By eliminating features with high collinearity, feature redundancy is reduced, and model stability is optimized.

(3) Sample Splitting: The dataset is divided into training and testing sets, used respectively for model training and performance validation. This ensures an objective evaluation of model performance while preventing overfitting, which could compromise the model's generalization capability.

(4) Algorithm Selection: Random Forest, SVM, Logistic Regression, and XGBoost were chosen to address challenges posed by different data characteristics. These algorithms ensure the ability to capture nonlinear relationships while balancing model accuracy and interpretability.

(5) 10-Fold Cross-Validation: Through multiple random groupings and validations, this method comprehensively evaluates model stability and reduces performance bias caused by the randomness of data splitting.

(6) Model Evaluation: The performance of the model is comprehensively evaluated using metrics such as AUC, ROC curves, F1 score, precision, and recall to select the best predictive model.

(7) Model Interpretation: SHAP is used to quantify the contribution of each features to the prediction outcome, enhancing the model's transparency and interpretability.

The overall process is illustrated in Figure 1.1.

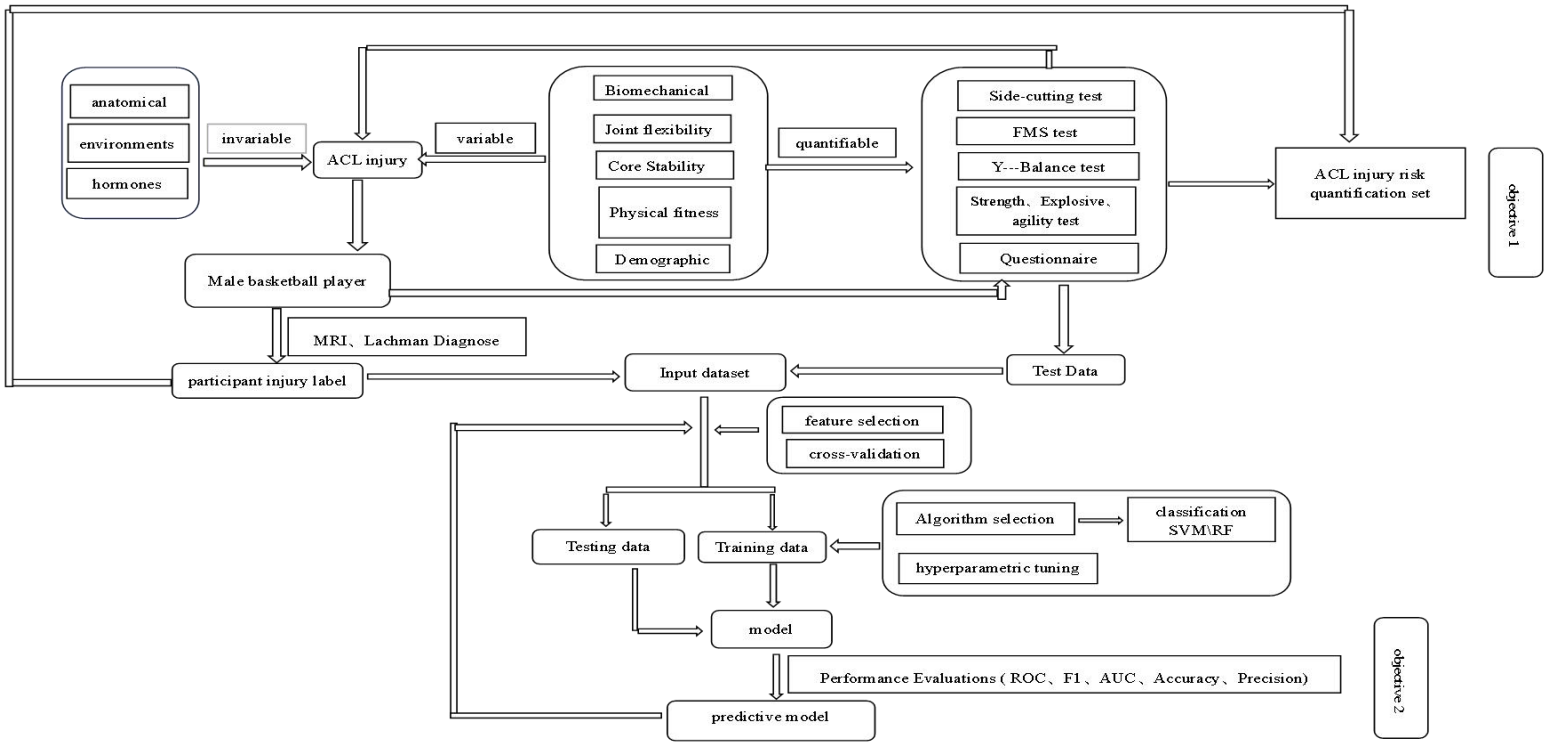

Figure 1.1 Theoretical framework of the study

1.4 Problem statement

About 70% of ACL injury are non-contact injuries (Claudino et al., 2019). Non-contact ACL injury is influenced by various risk factors, including anatomical structure, hormonal levels, and environmental conditions, many of which are unmodifiable. However, modifiable features of ACL injury such as biomechanical and neuromuscular characteristics can be improved through training, offering potential for injury prevention measures. Existing research has yet to elucidate the specific relationship between these modifiable features and ACL injury risk in male basketball players. Additionally,the anatomical features of ACL injury are highly variable in terms of gender and measurement of parameters (Jagadeesh et al., 2021). Therefore, in order to draw more accurate conclusions, the current research limits the research participants to male basketball athletes.

Although machine learning models have shown promise in predicting sports injuries (Johnson et al., 2019; López-Valenciano et al., 2018; Rommers et al., 2020), several limitations remain in ACL injury prediction. First, the included risk factors in existing studies vary significantly, leading to inconsistent prediction results and limited clinical applicability. Second, the issue of class imbalance is often overlooked, causing models to favor the majority class and reducing predictive performance for the minority class (i.e., injury cases). Third, the lack of interpretability in machine learning models results in insufficient transparency between features and prediction outcomes.

Therefore, this study optimizes the modeling strategy, ensures data validity, reveals key ACL injury factors, and enhances model interpretability.

1.5 Research Question

This study aims to investigate key issues in ACL injury prediction by addressing the following questions

(1) Which specific metrics are significantly associated with ACL injury risk in male basketball players?

(2) How valid the machine learning algorithms in predicting ACL injury among male basketball player?

(3) How do algorithms like Random Forest, XGBoost, Support Vector Machine (SVM), and Logistic Regression perform in predicting ACL injury?

(4) Can interpretability tools (e.g., SHAP) effectively identify key features associated with ACL injury?

1.6 Research objectives

1.6.1 Overall Objective

To develop and validate a machine learning framework that integrates biomechanical, physical, and demographic risk factors to predict ACL injury risk in male basketball players, identify the most effective predictive algorithm, and utilize SHAP for model interpretability to uncover key risk factors for ACL injury prevention.

1.6.2 Specific Objectives

(1) To evaluate the relationship of lower limb biomechanics, joint flexibility, core stability, physical fitness, athletes' profile features and incidence of ACL injury in male basketball players.

(2) To validate machine learning algorithm in predicting ACL injury among male basketball players

(3) To compare the performance of machine learning algorithms including Random Forest, SVM, Logistic Regression, and XGBoost and determine the best algorithm for ACL injury prediction.

(4) To perform visualization analysis of machine learning models, by evaluating the rankings of features across different algorithms using SHAP.

1.7 Research hypotheses

The following are the study hypotheses

Objective 1:

Ho: There are no significant relationships between lower limb biomechanics, joint flexibility, core stability, physical fitness, athletes' profile features and incidence of ACL injury

H_A: There are significant relationships between lower limb biomechanics, joint flexibility, core stability, physical fitness, athletes' profile features and incidence of ACL injury

Objective 2:

Ho: The predictive model is not valid in distinguishing between male basketball players with and without ACL injury.

Ha: The predictive model is valid in distinguishing between male basketball players with and without ACL injury.

Objective 3:

H_0: There is no significant difference in the classification performance of Random Forest, SVM, Logistic Regression, and XGBoost algorithms for ACL injury prediction.

Ha: There is a significant difference in the classification performance of Random Forest, SVM, Logistic Regression, and XGBoost algorithms for ACL injury prediction.

Objective 4:

H_0: SHAP analysis indicates that feature importance does not converge across different algorithms, making it impossible to identify consistent key features for ACL injury.

Ha: SHAP analysis indicates that feature importance converges across different algorithms, enabling the identification of consistent key features for ACL injury.

1.8 Significance of the study

This study holds significant theoretical and practical implications. First, by identifying key risk factors associated with ACL injury in male basketball players, this study helps athletes implement preventive measures, reduce injury rates, and prolong their careers. Second, by optimizing modeling strategies and incorporating various techniques—such as VIF for features screening, handling imbalanced samples, and cross-validation—it enhances the model's predictive performance and robustness, offering valuable insights for modeling in complex data environments. Gaussian noise combined with SMOTE is introduced to address sample imbalance,

and cross-validation is used to evaluate model stability. Additionally, multiple machine learning algorithms are employed, each offering unique advantages in injury prediction, providing multidimensional support for sports injury risk assessment. Random Forest integrates multiple decision trees, offering robust nonlinear modeling capabilities to capture feature interactions in complex data. Additionally, its built-in feature of importance evaluation mechanism supports the identification of key injury risk factors (Briand et al., 2022). SVM excels in high-dimensional feature modeling, capturing nonlinear relationships through kernel functions. Its margin-maximizing property effectively prevents overfitting, making it suitable for injury data analysis with limited sample sizes (Miawarni et al., 2022). Logistic Regression is renowned for its simplicity and interpretability, clearly demonstrating the influence of risk factors and enabling quantitative assessment of injury risk through probability outputs (Stylianou et al., 2015). XGBoost excels in complex feature modeling with its strong predictive performance and efficiency, enhancing model stability and generalization through iterative optimization and regularization mechanisms (Priscilla et al., 2020). Finally, SHAP visualization tools are utilized to rank features across different algorithms, identify key features that influencing ACL injury risk and aid in the development of targeted prevention strategies.

1.9 Operational definition

This study establishes a series of operational definitions to ensure consistency in understanding key concepts, and methodology throughout the research, (Table 1.1).

Table 1.1 Operational definition

Terms	Operational definition
Machine Learning	The extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model (Murdoch et al., 2019). It provides a framework for predicting outcomes and identifying key features influencing ACL injury
SVM	Support Vector Machine is a universal learner. It can classify both linear and non-linear data, (Joachims, 1999). It serves as a universal learner and is particularly effective for high-dimensional datasets
Logistic regression	Logistic regression is a statistical model that quantifies the relationship between a categorical dependent variable (e.g., ACL injury status) and one or more independent features. It is widely used for binary classification tasks due to its simplicity and interpretability (Nick & Campbell, 2007).
XGboost	Extreme Gradient Boosting (XGBoost) is an improved algorithm based on Gradient Boosting Decision Trees, which combines individual learners together through the boosting technique to establish dependencies. Additionally, it can effectively construct boosting trees and run in parallel (Jiang et al., 2023).
Random forest	A random forest (RF) is an ensemble classifier and consisting of many Decision Trees similar to the way a forest is a collection of many trees (Breiman, 2001). It is particularly effective in handling non-linear data and identifying features importance
ACL injury	In this study, ACL injury refers specifically to non-contact anterior cruciate ligament injuries of varying severities, diagnosed using Magnetic Resonance Imaging (MRI) or confirmed by positive Lachman tests. This operational definition ensures consistency in identifying the target condition.

Table 1.1 Continued

Male basketball player	The study focuses on male college basketball players aged 18 and above who have participated in competitive basketball for at least three years. This criterion ensures the inclusion of athletes with adequate exposure to the physical demands and injury risks inherent in basketball.
Physical Function Features	Physical Function Features are defined as measurable features of an individual's physical capabilities that influence variations in injury risk. Key Physical Function Features include the Y-Balance Test (Left Leg Combined Score, Right Leg Combined Score, Double Leg Difference) and the overall Functional Movement Screen (FMS) score, which collectively assess stability, mobility, and movement symmetry.
Athlete's Profile Features	Athlete's profile features are defined as individual characteristics encompassing physical attributes, personal history, and lifestyle features that influence injury risk. These include measurable physical parameters (e.g., height, weight, age, and body composition), training background (e.g., sport level, years of training, and weekly training hours), functional roles (e.g., playing position), and health-related information (e.g., injury history, hereditary conditions, and medication use). Collectively, these features provide a comprehensive understanding of an athlete's predisposition to injuries, enabling tailored risk evaluation and the formulation of preventive measures.

Table 1.1 Continued

Physical Fitness features	Physical Fitness features represent the athletic abilities and performance metrics directly related to basketball activities that influence ACL injury risk. These features capture the strength, power, agility, and explosive performance of players. Examples include lane agility, one-repetition maximum (1RM) for squat and deadlift, relative strength metrics (e.g., relative squat and deadlift loads), peak maximum rate of force development (mRFD) during squat and deadlift, dominant single-leg hop distance, and vertical jump performance (e.g., squat jump, deep jump, countermovement jump). These measures provide insights into a player's neuromuscular capabilities and biomechanical efficiency, which are critical in assessing injury risk during basketball-specific movements.
Biomechanical features	Biomechanical features pertain to the kinematic, kinetic, and electromyographic parameters observed during unanticipated side-cutting movements that influence ACL injury risk These features involve dynamic movement patterns, joint angles, moments, and muscle activations that occur during high-risk maneuvers. Key features included the standardization of the center of mass (COM), ground reaction forces measured via force platforms, horizontal distances from the ankle joint to the center of pressure, and specific joint angles and moments . Additionally, electromyographic (EMG) activation levels of key muscles, including the rectus femoris, vastus medialis, vastus lateralis, biceps femoris (long and short head), medial and lateral gastrocnemius, are measured to evaluate neuromuscular control during these movements. These data allow for a detailed analysis of the biomechanical and neuromuscular mechanisms underlying ACL injury risk.

CHAPTER 2

LITERATURE REVIEW

2.1 Anterior cruciate ligament: structure and injury diagnosis

The human body's largest and most complex weight-bearing joint is the knee joint which comprises the lower end of the femur, the upper end of the tibia, and the patella (Figure 2.1) (Hughes & Watkins, 2006). In addition, the knee joint is supported by a combination of ligaments, muscles, and tendons that work together to provide strength and flexibility. It enables us to do various physical activities like walking, running, bending, rotating, squatting, and complex movements (Xu et al., 2007). The anterior cruciate (ACL) and posterior cruciate (PCL) ligaments have a hinge shape and are attached between the intercondylar fossa of the femur and the intercondylar eminence of the tibia. These ligaments play a crucial role in maintaining proper knee joint function. Any injury to the ligaments can destabilize the knee joint, hinder physical activity, and lead to secondary osteoarthritis (Kraeutler et al., 2017; Lohmander et al., 2007).

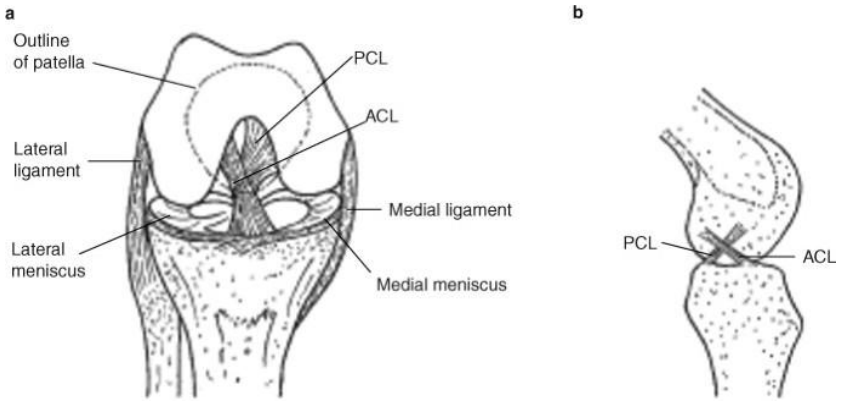

Figure 2.1 Anatomy of the knee joint. (Adopted with permission from Hughes & Watkins, 2006).

The ACL is a crucial component for the knee joint stability by controlling the excessive movement of the tibia and femur (Escamilla et al., 2012). It is a strong band made of connective tissue and collagenous fibres that originate from the anteromedial aspect of the intercondylar region of the tibial plateau and extends posteromedially to attach to the lateral femoral condyle (Gupta et al., 2019). The ACL is formed by anteromedial bundle and posterolateral bundle. ACL injury in sports occur due to excessive motion involving flexion, extension, and twisting of the knee (Evans & Nielson, 2023). Recovery typically requires at least 6 months of surgical reconstruction and rehabilitation (Kruse et al., 2012). It is estimated that the yearly expense for extended care following an ACL injury is approximately $2.8 billion (Mather et al., 2013).

At present, physicians rely on a combination of a patient's medical history, physical examination, and Magnetic Resonance Imaging (MRI) (Musahl et al., 2018) to accurately diagnose ACL injury (Bai et al., 2022). In cases of acute ACL injury, patients may hear a tearing sound or feel an abnormality in the joint, followed by intense knee pain, limited mobility, and swelling. Other signs include persistent swelling in the knee, a sensitive knee area, or unstable knee movements while

engaging in activities that involve twisting or jumping (Peat et al., 2017). When physicians suspect that an athlete has an ACL injury, they commonly use the Lachman test and pivot shift test to evaluate the severity of the injury (Bai et al., 2022). In addition, physicians frequently use a combination of MRI techniques to diagnose ACL tears. MRI offers several benefits such as multi-directional imaging, good contrast for soft tissues, and high spatial resolution. It also displays exceptional sensitivity and specificity in detecting ACL tears (Phelan et al., 2016).

The classification of ACL tears is based on the extent of the ligament damage, with the following being the commonly used medical standards (Dar et al., 2022): Grade I - minor strain of the ligament without significant tearing, knee remains stable, mild pain and swelling, Lachman Test shows minimal movement (less than 5 mm); Grade II - partial tear of the ligament, affecting knee stability. The patient may experience knee instability accompanied by moderate pain and swelling. Lachman Test shows noticeable movement (5-10 mm); Grade III - complete ligament tear, resulting in knee instability. This injury grade is usually associated with severe pain, swelling, and inability of the knee to bear weight. while the Lachman Test shows a forward movement of more than 10 mm. The current study focuses on all grades of ACL injury with a positive Lachman Test and a comprehensive assessment of the injury using MRI.

2.2 ACL injury mechanisms and risk factors

In high-intensity sports like basketball and soccer, movements changing direction, jumping, landing, and sudden stops can increase the risk of ACL injury (Leppänen et al., 2017). ACL injury occur when there is instability (i.e., abnormal

motion) in the tibio-femoral joint, which causes an overload on the ACL (Hughes & Watkins, 2006). The tibio-femoral joint, commonly known as the knee joint, exhibits motion across three anatomical planes: the sagittal, frontal, and transverse planes, reflecting its complexity and versatility (Figure 2.2) (Kwong, 2008). This motion is characterized by six degrees of freedom, encompassing three rotational movements and three translational movements. The rotational movements include flexion and extension in the sagittal plane, internal and external rotation in the transverse plane, and abduction and adduction in the frontal plane. The translational movements consist of anterior-posterior translation (i.e., sliding forward and backward), medial-lateral translation (i.e., shifting side to side), and proximal-distal translation (i.e., compression and distraction along the joint axis). Together, these degrees of freedom enable the knee joint to perform its essential functions in dynamic activities while maintaining joint stability and mobility. ACL injury is often caused by the knee being outside of its normal range of motion in all directions. This can result in excessive strain on the joint, ligaments, and cartilage. Studies have shown that the ACL experiences greater stress when subjected to complex and multidirectional loading, compared to loading in a single direction (Berns et al., 1992; Markolf et al., 1995; Miles et al., 2022). Athletes who play basketball are at a higher risk of experiencing ACL injury due to the complex and varied nature of the movements involved in the game.

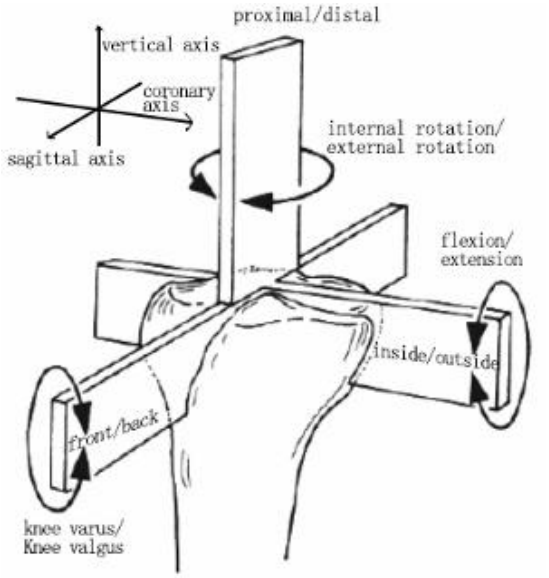

Figure 2.2 Six degrees of freedom (i.e., three rotational, three translational) of the knee joint motion model (Adopted with permission from Kwong, 2008)

Approximately 70% of ACL injury are non-contact in nature, making them the primary focus of this study (Claudino et al., 2019). Non-contact injuries occur without any direct physical contact between the athlete and an external force, such as another player or object. In contrast, contact injuries involve external forces, such as a collision with another player or a direct blow to the knee (Zhou, 2018). Non-contact ACL injury is often multifactorial and the risk factors can be categorized into modifiable and non-modifiable features (Zhou, 2018). Non-modifiable risk factors include, anatomical structures (Simon et al., 2010), hormones (Hewett et al., 2007) and environmental features (Orchard & Powell, 2003), which are often difficult or impossible to change. Modifiable risk factors include biomechanics and neuromuscular control associated with ACL injury (Zhang et al., 2019). Accurately identifying the modifiable risk factors, is crucial to develop effective injury prevention programs to reduce ACL injury incidence.

2.2.1 Anatomical risk factors

When the knee is in an overly extended position in the sagittal plane, all the fibres of the ACL are stretched. The base of the intercondylar fossa then pulls it into a curved chord, which serves as the primary resistance to hyperextension (Wang et al., 2019). When the ACL ligament is stretched, it becomes vulnerable to damage from shear stresses.Therefore, athletes with larger quadriceps angles (also known as Q-Angle) and flat foot may experience an increase in knee valgus and a decrease in lower limb stability which can greatly increase the risk of ACL injury when external forces impact the knee (Vacek et al., 2016). Combined features of the medial tibial depth (MTD) and lateral posterior tibial slope (LPTS) are important anatomical factors in assessing the risk of ACL injury (Misir et al., 2022). It was shown that combined shallow MTD Shallow and decreased LPTS is a risk factor for ACL injury in male, while combined shallow MTD and increased LPTS is a risk factor for ACL injury for female (Hashemi et al., 2010). There are certain sex-specific anatomical factors, such as Tibial Slope (TS) and lateral femoral condyle ratio (LFCR), which pose a risk solely to female (Barnum et al., 2021; Beynnon et al., 2014; Jeon et al., 2022). On the other hand, the ACL size and the shape of the intercondylar notch are strongly associated with an ACL injury risk in male (Whitney et al., 2014; Huang et al.,2020).

Additionally, the posterior slope angle of medial tibial plateau is a contributing factor to the risk of ACL injury (Matas et al., 2021). The articular surface of the medial tibial plateau is not perpendicular to the longitudinal axis of the tibia, but rather has a downward sloping angle. When an athlete abruptly changes direction and lands on one foot, it creates pressure on the knee and causes the quadriceps muscles