RUNNING-RELATED INJURY CLASSIFICATION FOR PROFESSIONAL RUNNERS

BY:

DARWINESWARAN A/L RAJA LINGAM

(Matrix no.: 137803)

Supervisor:

Associate Professor Dr. Loh Wei Ping

This dissertation is submitted to

Universiti Sains Malaysia

As partial fulfilment of the requirement to graduate with honors degree in

BACHELOR OF ENGINEERING (MECHANICAL ENGINEERING)

School of Mechanical Engineering

Engineering Campus

Universiti Sains Malaysia

12th July 2021

DECLARATION

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

Signed…… Darwin …… (DARWINESWARAN A/L RAJA LINGAM)

Date……12 / 7 / 2021……

Statement 1: This journal is the result of my own investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. Bibliography/ references are appended.

Signed…… Darwin …… (DARWINESWARAN A/L RAJA LINGAM)

Date……12 / 7 / 2021……

Statement 2: I hereby give consent for my journal, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed…… Darwin …… (DARWINESWARAN A/L RAJA LINGAM)

Date……12 / 7 / 2021……

ACKNOWLEDGEMENT

This thesis would not have been possible without the help, support and guidance of many people. First and foremost, it is with a great deal of respect and gratitude that I acknowledge my Final Year Project (FYP) supervisor, Associate Professor Dr. Loh Wei Ping. Her consistent support and ideas had motivated me to carry out this research to my utmost ability. During the challenging Covid-19 pandemic, her aid in online discussions and giving valuable feedbacks whenever I approach her have been very helpful.

I would like to show my sincere appreciation to the entirety of technical staff and lecturers of School of Mechanical Engineering that have aided me throughout my 4-years undergraduate studies. Their support, guidance, and encouragement have enhanced my skills, knowledge, and proficiency in various ways.

I would also like to thank my friends who have motivated me to keep going throughout this difficult period. They have helped me financially and emotionally all along my journey at the university.

Lastly, not forgetting my parents, who I owe an eternal debt to, for being my supporting pillars throughout my life. The opportunities they have presented to me has allowed me to discover my purpose in life. I express sincere thanks and deep appreciation for all your supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AT	Achilles Tendinopathy
BMI	Body Mass Index
CSV	Comma-Separated Values
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
IBk	Instance Based Learner
ITBS	Iliotibial Band Syndrome
KNN	K-Nearest Neighbors
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
MS Excel	Microsoft Excel
MTSS	Medial Tibial Stress Syndrome
PFPS	Patellofemoral Pain Syndrome
ROC	Receiver Operating Characteristics

RRI	Running-Related Injury
SMOTE	Synthetic Minority Oversampling Technique
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
WEKA	Waikato Environment for Knowledge Analysis

ABSTRAK

Berlari adalah satu bentuk aktiviti fizikal yang sihat, tetapi boleh terdedah kepada kecederaan jika terlibat secara berlebihan atau dilakukan dengan postur yang tidak betul. Kajian terdahulu telah mempertimbangkan faktor risiko kecederaan berkaitan berlari (RRI) terhad dan mempunyai asal usul multifaktorial.

Walau bagaimanapun, tidak banyak yang dibincangkan mengenai parameter ramalan yang harus dipertimbangkan ketika mengkaji jenis risiko kecederaan yang mungkin mempengaruhi pelari tertentu. Kajian ini bertujuan untuk mengkaji kualiti kumpulan data berkaitan aktiviti berlari untuk membuat analisis klasifikasi kecederaan-kecederaan yang boleh dipercayai dengan WEKA dan juga untuk mewujudkan model klasifikasi yang sesuai tentang RRI untuk pelari profesional.

Data daripada 74 pelari profesional dikumpulkan dari repositori Kaggle. Kumpulan data ini terdiri daripada kelas yang cedera dan tidak mengalami kecederaan yang diukur dengan atribut data (jumlah hari rehat, jumlah km Z3-Z4-Z5-T1-T2, jumlah km Z3-4, jumlah km Z5-T1-T2, jumlah latihan alternatif, jumlah latihan kekuatan, rata-rata latihan rutin, rata-rata kejayaan latihan, dan rata-rata pemulihan). Analisis klasifikasi dilakukan pada data kajian menggunakan algoritma BayesNet, RandomForest, J48, RandomTree, REPTree, dan IBk dalam toolkit WEKA. Set data RRI telah diproses terlebih dahulu untuk menyaring nilai luaran dan nilai ekstrem serta atribut data yang tidak relevan sebelum klasifikasi. Hasil kajian menunjukkan bahawa tiga algoritma pengklasifikasi terbaik dengan ketepatan tertinggi untuk mengklasifikasikan pelari ke dalam kategori tidak cedera dan cedera adalah BayesNet (98.6457%), RandomForest (98.0107%), dan (tidak terpotong) J48 (97.1002%). Penyelidikan ini merupakan langkah maju dalam meramalkan kemungkinan RRI pada pelari profesional menggunakan pendekatan perlombongan data.

ABSTRACT

Running being a form of healthy physical activity which is prone to injuries if performed excessively or with incorrect posture. Previous studies have considered risk factors of running-related injuries (RRI) to be limited and have multifactorial origins.

However, little is discussed on prediction parameters to be considered when studying the type of potential injury risks that may affect a particular runner. This study aims to investigate the qualities of RRI dataset for reliable running-injury classification analysis with WEKA and also to establish an appropriate classification model for RRI in professional runners.

The data from 74 professional runners were collected from Kaggle repository. This dataset consisted of injured and uninjured classes measured by data attributes (nr. rest days, total km Z3-Z4-Z5-T1-T2, total km Z3-4, total km Z5-T1-T2, total hours alternative training, nr. strength trainings, avg exertion, avg training success, and avg recovery). Classification analyses were performed on study data using BayesNet, RandomForest, J48, RandomTree, REPTree, and IBk algorithms in WEKA toolkit. The RRI dataset was pre-processed to filter outliers and extreme values as well as irrelevant data attributes prior to the classification. Findings revealed that three best classifier algorithms with the highest accuracies to classify runners into the category of uninjured and injured are BayesNet (98.6457%), RandomForest (98.0107%), and (unpruned) J48 (97.1002%). This research is a step forward in predicting a probable RRI in professional runners using a data mining approach.

Chapter 1 INTRODUCTION

1.1 Overview

This chapter introduces about running-related injuries and the brief background covering this issue. The problem statement of past studies on running-related injuries were presented. The objectives of the research work were listed along with the study scope as well as the outline of the entire thesis.

1.2 Background

Running is one of the most popular physical activities around the world to achieve or maintain better physical health (van Poppel et al., 2020). Running is increasingly popular and is associated with longevity. However, distance running is associated with high incidence of lower limb injuries (Ellison et al., 2020 and Mann et al., 2016). This may be due to the altered state of foot strike pattern from a predominantly forefoot strike (FFS) landing to a predominantly rearfoot strike (RFS) landing (Davis et al., 2017).

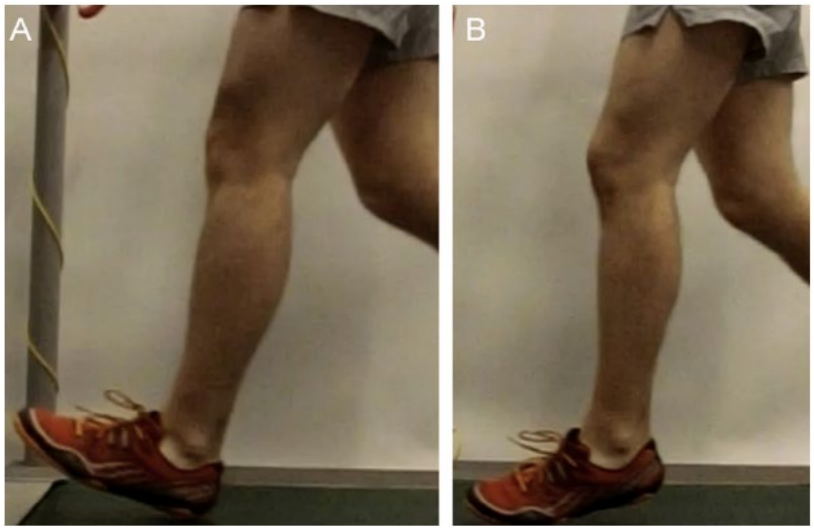

Figure 1.1 Lower extremity alignment at foot strike: (A) Rearfoot strike, (B) Forefoot strike (Davis et al., 2017)

The use of footwear to run, barefoot running, and the type of runners (age, gender) that particularly runs on hard surfaces like road are some of the parameters commonly studied (Davis et al., 2017). Other factors also include the processed food consumed, the polluted air breathed in, and the relative lack of activity people engaged in (Davis et al.,

2017). These were the common variables which may affect the runner's lower extremities to potentially develop a running-related injury. Most works of gathering information on running-related injuries were from self-reported surveys with a small representative sample and they were dedicated to improving running technique and identifying potential predictors of injury (Warne et al., 2021). For instance, in the field of sport science, a study conducted by (Hamil & Gruber, 2017) on the changes of foot strike pattern from a rearfoot to a mid- or forefoot strike suggest that there are no obvious benefits to the majority of runners. In fact, change in foot strike may result in stressing the tissue when running. Furthermore, when associated with the medical field, studies on previous injuries were conducted by Hespanhol Junior et al., (2013) to predict running-related injuries to enable newer training methods as well as protective measures for runners in general.

Real time tracking of runner's performance data are usually collected by using wearable technology like heart rate monitoring device, acceleration tracking, collision impact tracking, and sweat analysis (da Araujo et al., 2015). Some real-life application which uses running data includes fitness mobile application (Runkeeper, Nike Run Club, Adidas Running App by Runtastic), wearable fitbit, and sport science as well as medical usage.

Experiments have been conducted according to protocols with manual input of runner's experimental performance, to using a subject-specific mathematical model which includes an accurate metatarsal geometry (Ellison et al., 2020). The number of selective participants to include was predefined with written consent. Prior to data analysis, the data will undergo cleaning, extraction and clustering into categories for informative knowledge discovery.

Apart from experimental study, running data were also obtained from the public data domain. For example, some studies were conducted using a self-report online survey form where respondents manually fill in demographics like age, gender and previous injuries (Warne et al., 2021). In addition, some related public databases were made accessible at Figshare, EasyBib.com, and ReadCube Papers (Mousavi et al., 2021). These data are openly available to the public to benefit future researches on running-related injuries.

1.3 Problem Statement

The data obtained from historical running records are insufficient to give an exact indicator of potential running injury. This is due to every individual runner has different demographic factors like gender, BMI, lifestyle, and feet size. When these variables are taken into account, running performances may vary and differ from the average generic performance model. It is noteworthy that the gender and age of runners tend to have varying impact on potential running-related injury. Variables like foot strike pattern and running surface influences are also important to determine running related injury levels to occur. However, these aspects are still lacking in existing studies. Researches often focused on exploring new methods to interpret the effects of running on runner's joints and performance (Warne et al., 2021). Not much work is done to classify running injury. Such aspects contribute a greater challenge to interpret, extract, clean and develop a predictive classification model for running injuries.

Many studies focused on the previous injuries sustained by runners like in Hespanhol Junior et al., (2013) and Malisoux et al., (2015). They discussed the risk factors that previous injuries pose as potential running injury for the future. Meanwhile, no other studies attempted to predict parameters to be considered when studying the type of potential injury risks that may affect a particular runner. Data mining analysis can be used to analyze and classify if a subject is prone to developing an injury based on the work done and the effort exertion. Therefore, a study on classifying runners based on type of training will be conducted with data mining technique, so as to understand a balance need of non-injured and injured running under various factors, allowing a new breakthrough in the aspect of predicting running injury risks.

1.4 Objectives

The focus of this study is:

- To investigate the qualities of dataset for reliable running-injury classification analysis
- To establish an appropriate classification model for running-related injury in professional runners

1.5 Scope of Work

The study includes a data mining approach in classifying potential running-related injury case study. This case study involves a publicly available dataset consisting of 74 professional and experienced Dutch runners with varying gender, age, resting period, training period and running route.

The study utilizes the open-source data mining software WEKA for data cleaning, filtering, pre-processing and classification as well as verifying and building the classification model. This study focuses on binary classification of injured and uninjured classes of runners associated with the quantitative data attributes: nr. rest days, total km Z3-Z4-Z5-T1-T2, total km Z5-T1-T2, total hours alternative training, nr. strength trainings, avg exertion, avg training success, and avg recovery.

1.6 Thesis Outline

This thesis is divided into 5 main chapters. Chapter 1 is the introduction to the study. It includes project background, the problem statement, the study objectives, and the study scope.

In Chapter 2, a review of published information related to running-related injury from the past studies are discussed. The related articles are explored from the issues encountered and limitation in the running analysis.

In Chapter 3, the methodology of the entire research study is explained. This chapter explains the sequential processes flow beginning from the dataset used followed by data mining analyses at data pre-processing and classification stages. The methods and techniques applied in this study are discussed.

Chapter 4 presents the results generated from mining the running-related injury. Findings from the qualitative enhancement of the dataset are presented. Besides, the classification results are also discussed and compared with the previous studies.

Last but not least, Chapter 5 concludes the study by providing an overall view of the main outputs of running-related injury classification. The chapter also outlines how the objectives of this study were achieved, contributions of this study as well as suggestions for the future work.

Chapter 2 LITERATURE REVIEW

2.1 Overview

This chapter presents the state-of-the-art review on past research presented on running-related injuries and factors associated in developing a running-related injury. A total of 50 related articles published in year 1991 to 2021 were consulted. The search strategy for these 50 journal articles is shown in Figure 2.1. The type of journals and the origin of journal are attached at the Appendix.

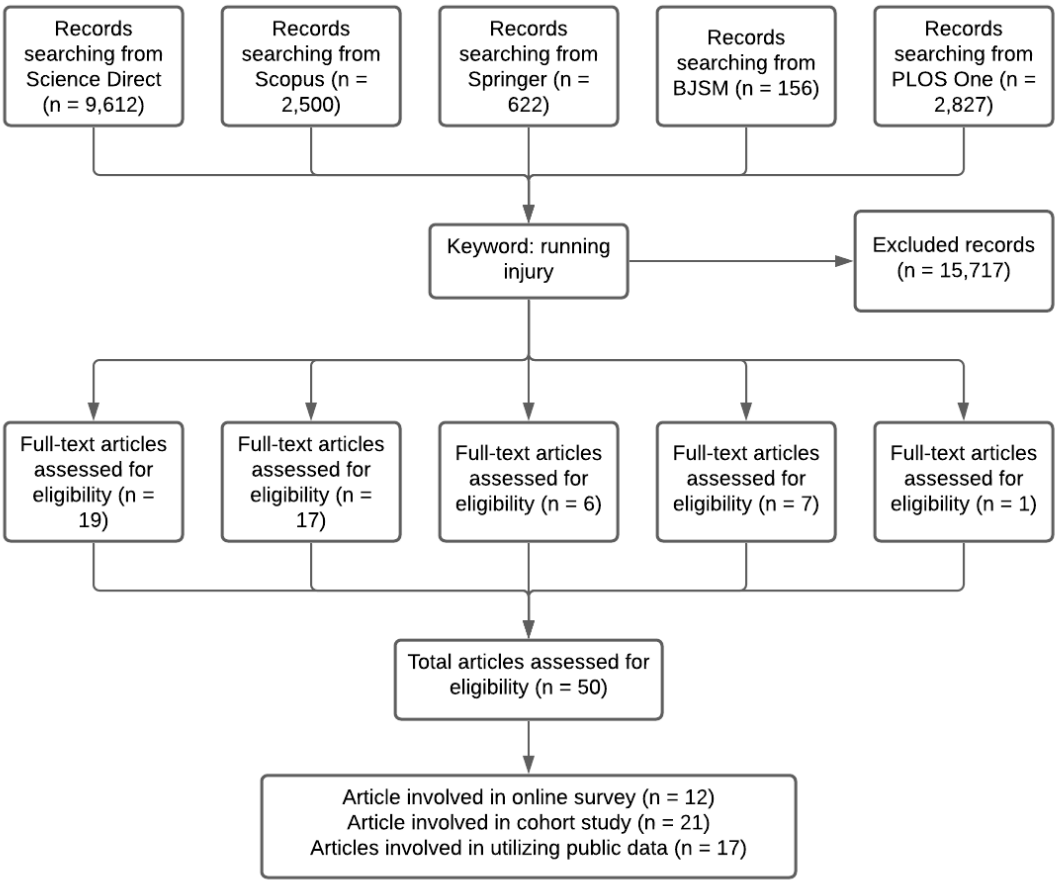

Figure 2.1 Schematic diagram for literature review search strategy

2.2 Definition of RRI

The incidence of RRI refers to the frequency of running-related injury occurring. Biological structures like the tendons, bones, ligaments, and muscles adapt positively if they are exposed to repeated stress. This only applies if the stress exposed is within the

dilative limit, followed by adequate rest in between the applied forces. If an applied stress exceeded the tolerance limit, a scarring injury may occur. RRI can develop when repetitive stressors are applied slightly below or at the dilative limit, generating microtrauma when there are insufficient rest periods for rejuvenation, according to Hreljac and Ferber (2006).

Nevertheless, at present, there does not exist a standard definition for RRI, which causes difficulty comparing the incidence of running-related injuries. For example, a study done by Warne et al. (2021) defined a running injury as "any discomfort or pain that caused a runner to miss at least two days of training, and/or required treatment from a medical-experts." Similarly, Buist et al. (2008) defined running injury as "any musculoskeletal complaint of the lower extremity or back restricting running for at least one week." The same author, Buist et al. (2010) have later reduced the duration of running deficiency to only one day. Besides, Taunton et al. (2003), grouped runners as injured only if runners experienced pain only on completion of a race or run. To top it off, the definition of running-related injuries was also separated into different locations for specific injury.

2.3 Incidence of RRI

The rate of reported running-related injuries varies by author and is heavily influenced by the study methodology. For example, Goss and Gross (2012) found that the average runner who consistently trained and occasionally competed had a yearly incidence rate of 37-56 %. However, a review by van Gent et al. (2007) reported injury rates ranged from 19.4-79.3%.

The rate of injuries reported has been more consistent across different time frames as a result of observing the outcomes of training programs reported in the past. For an 8-week training period, the rate of injury was 25.9% (Buist et al., 2010). Meanwhile, in a 12-week training period, 31% of novice runners sustained a running injury (Hespanhol et al., 2013). Furthermore, there was a 29.5 % injury rate in a 13-week running program designed to reduce injury (Taunton et al., 2003). The injury rate was 79 % and 59 % per 1000 hours of running time in a six-month follow-up (Lun et al., 2004). Buist et al. (2010) revealed an injury rate of 30.1 injury counts per 1000 hours of running. Murphy et al. (2003) reported the injury rate in athletic exposure (AE) to be calculated at 17.0 injuries per 1000 AEs.

Varied studies utilize different training durations, data collection methods, and phrasing for the incidence rates by percentage per hour as well as injury counts, making it difficult to compare reported injury rates from the literature. Goss and Gross (2012) stated that "the incidence of sport injuries depends on the methods used to count injuries for example either in prospective or retrospective methods. Those are used to establish the population at risk and on the representativeness of the sample". He claimed that the definitions of sports injury, sports injury incidence, and sports participation all have a role in the acute and overuse sports injury problem. Some injuries are traumatic such like, ankle sprains. However, most injuries are overuse in nature (Mercer et al., 2015; Hreljac and Ferber, 2006; Goss and Gross (2012)).

The most commonly reported RRI is Patellofemoral Pain Syndrome (PFPS). PFPS is a medical condition that causes pain under or around the kneecap. PFPS can occur in one or both knees and can affect both children and adults. The second most frequently mentioned injury is Iliotibial Band Friction Syndrome (ITBS). ITBS is a type of knee injury that causes discomfort and/or tenderness on probing of the lateral aspect of the knee, above the joint line and below the lateral femoral epicondyle. Other common injuries are Medial Tibial Stress Syndrome (MTSS) and Achilles Tendinopathy (AT) (Yong et al., 2020; Tokinoya et al., 2020; Bramah et al., 2018; Yao et al., 2021; Messier et al., 2018; Kozinc & Šarabon, 2017; Davis et al., 2017; Samaan et al., 2014; Lorimer & Hume, 2014; Yukawa et al., 2013; Pérez-Gómez, 2020; Taunton et al., 2002; Yeung and Yeung, 2001). MTSS refers to an overuse injury or repetitive-stress injury of the shin area. On the other hand, AT is a common overuse injury produced by excessive compression and recurrent energy storage and release. AT can produce a sudden injury or, in the worst-case scenario, an Achilles tendon rupture.

2.4 RRI by Bodily Location

The bodily location distribution associated with running-related injuries have been established throughout the literatures. Majority authors are in agreement with each other that the most frequent injury locations are at the lower extremity that include the foot, ankle, knee and hip (Warne et al., 2021; Viljoen et al., 2021; van Poppel et al., 2020; Pérez-Gómez, 2020; Hreljac, 2005; Taunton et al., 2002, 2003; Paquette et al., 2020). Figure 2.2 displays

the frequent locations for potential running-related injury to occur and the percentage distribution count running-related injury reported in Warne et al. (2021).

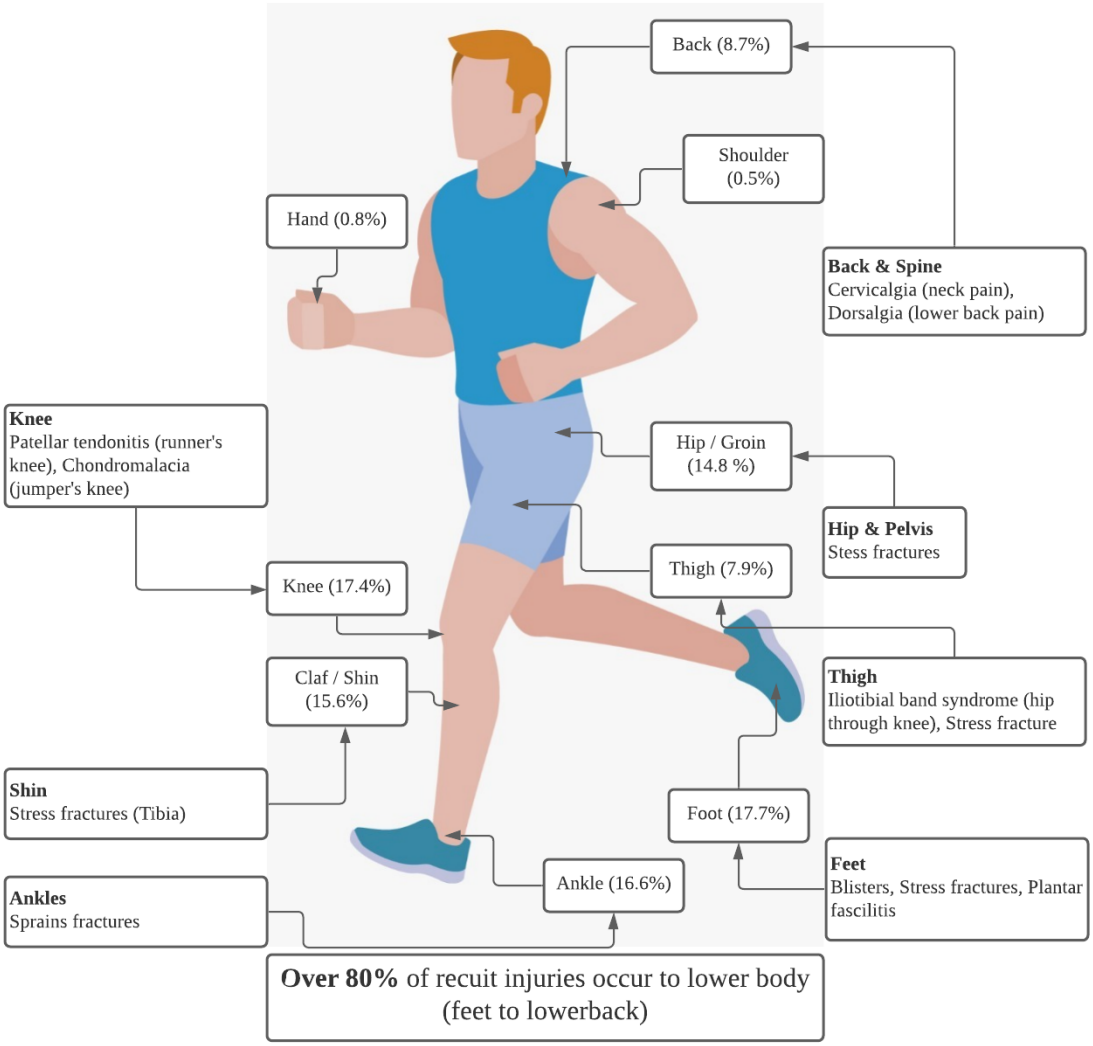

Figure 2.2 The anatomical location and distribution of individual injuries reported as a percentage and types of RRI that often occur

With reference to Figure 2.2 the lower extremity (shin, ankle, foot, knee, hip, and thigh) showed 90% of the injuries. This is due to the lower extremity often exposed to surrounding when running. On the other hand, the upper extremity (hand, back, and shoulder) only indicated a total of 10% on the injury in Figure 2.2. Apparently, these locations are less likely to develop a RRI while running.

2.5 Risk Factors of RRI

There are many risk factors that resulted in RRI. These factors can be divided into internal and external risk factors. The former can be segregated by anthropometry, anatomy, biomechanical while the latter can be detailed into training error and equipment used.

2.5.1 Internal Risk Factors

Internal risk factors also known as intrinsic risk factors are aspects that are cause within the runners. The internal risk factors cover the anthropometry, anatomy, and biomechanical aspects of the body (Figure 2.3).

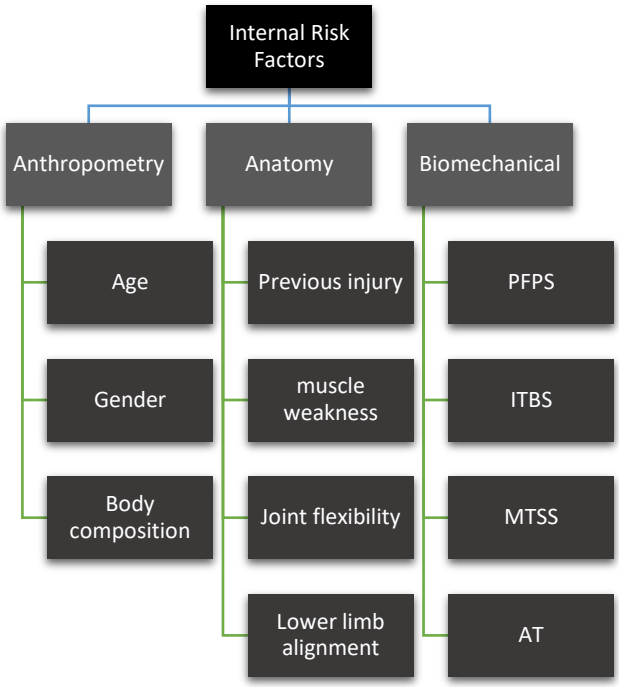

Figure 2.3 Internal risk factors

Anthropometry is the systematic measurement of the physical properties of the human body (Taunton et al. 2002). For this study we are interested in the age, the gender, and body composition. Taunton et al. (2002) analyzed the influence of age in 2002 patients suffering from running-related injuries. The study revealed that younger runners tend to be at higher risk of developing PFPS compared to old runners. As for gender, most literature agreed that no difference have been discovered in the injury rate between male and female runners (Saragiotto et al., 2014; Francis et al., 2019; Edwards et al., 2018). Meanwhile, for

body composition, studies often express body mass in kg/(height)^2, also known as body mass index (BMI). A greater BMI exerts extra stress on the body and thus a heavier runner is prone to higher risk of injuries (Buist et al., 2010).

Previous body damage, muscle weakness, joint flexibility, and lower limb alignment are all anatomical considerations to consider. A previous lower-extremity injury that has not fully healed is a possible risk factor for future running injuries. According to Taunton et al. (2003), half of the injured participants in his study had previously sustained injuries at the same anatomical region. In addition, 42% of the study participants were not fully recovered when they began the program. Weak muscles, specifically muscular imbalances, have been suggested as a possible cause of injury in runners. (Bender et al., 2019; Kozinc & Šarabon, 2017; Moffit et al., 2020). Joint flexibility is the range of motion of the muscle and cannot be mistaken for stretching. A lack of joint flexibility has been linked to an increase in muscle stiffness, which could result in additional stress on the corresponding joint (Yeung & Yeung, 2001). Besides, lower limb alignment is an alignment abnormality of various sorts and it is greatly suspected to cause running-related injuries. Lun et al., (2004) classified static lower limb alignment issue may cause injury to a specific part of the lower limb.

Biomechanical risk factors that are most frequently presented in the researched literature were PFPS, ITBS, MTSS and AT. PFPS is the most common overuse running injury and commonly known as the "runner's knee" (Huber, 2009). A synchronous coupling of the lower extremities, connected to the shoe and ability to adjust to the surface, has been identified as a factor in PFPS in a recent biomechanical study by Van Gent et al., (2007). ITBS is the most prevalent overuse injury for lateral knee pain and the second most common cause of knee discomfort among regular runners (Taunton et al., 2002). Moffit et al., (2020) confirmed that runners with ITBS had larger peak hip-adduction angles than the uninjured control participants. MTSS, also known as Shin Splints or Shin Pain, was considered to be caused by repeated stress during weight-bearing exercises such as running. For instance, Buist et al., (2010) reported that female high school cross-country athletes with greater foot pronation measured by navicular drop had an increased risk of developing MTSS. Meanwhile, AT is a clinical condition of the overused tendon. Researches have investigated the causes of AT, but findings are often different from other studies, maybe

due to differing study. The majority of biomechanical risk factors studied by Lorimer and Hume (2014) yielded ambiguous results, which is likely owing to the complex nature of Achilles overuse injuries.

2.5.2 External Risk Factors

External risk factors are external stresses exerted on the human body and for the many cases modifiable. These factors are categorized into training error and equipment (Figure 2.4).

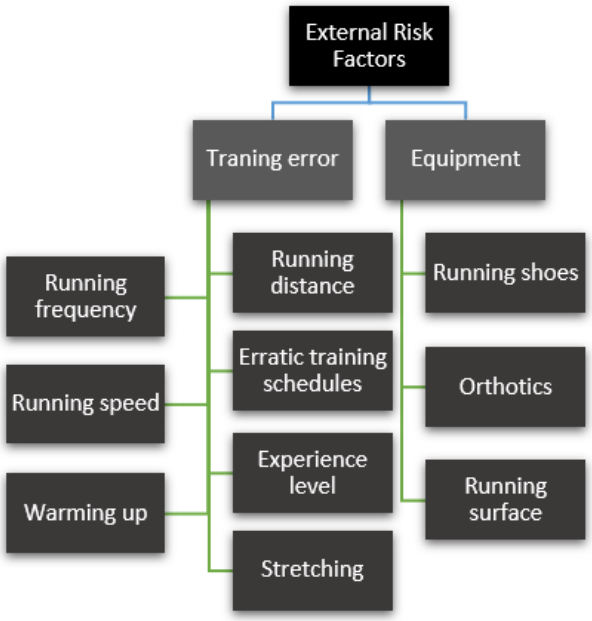

Figure 2.4 External risk factors

The training program experienced by a runner greatly impacts the development of an injury. In this aspect, the external factors involve training errors in running distance, running frequency, erratic training schedules, running speed, experience level, warm up procedure and stretching (Taunton et al. 2002). Several researches have reported unsupervised running (distance, frequency, speed) as one of the risks of sustaining a RRI (Barros et al., 2021; Malisoux et al., 2015; Kozinc and Sarabon, 2017; Moffit et al., 2020; Yeung & Yeung, 2001). It is revealed that mileage is a common training error that leads to RRI (Malisoux et al., 2015; Kozinc and Sarabon, 2017; Moffit et al., 2020; Yeung & Yeung, 2001; van Gent et al., 2007). Kozinc and Sarabon, (2017) revealed that a reduce in running distance resulted in lower injury rates among military recruits in their research. On the

other hand, Yeung & Yeung (2001) found that a high training load (intensity, frequency, or running duration) increases the risk of injury, and that changing the training plan can minimize the risk of injury. Malisoux et al., (2015) concurred with Taunton et al., (2002) that running distance and frequency are inextricably linked. As a result, it's possible that an increase in RRI risk is linked to weekly running distance. It is the result of a rise in running frequency rather than a decrease in mean session distance. Changes in a training routine, particularly in the short term, have been addressed as a possible injury risk factor (Taunton et al. 2002). But there were only a limited studies that investigated training related injury outcomes (Junior et al., 2017). Experience in running is an important factor related to RRI. According to Saragiotto et al., (2014), around 45 percent of those with an average running experience of 5.5 years had previously suffered an RRI. Warming up is similarly useless in reducing injury risks when it focuses just on stretching, according to Yeung & Yeung (2001). As a result, as the body warms up, as the temperature rises. The usefulness of stretching exercises and insoles in preventing lower extremity soft tissue injuries remained uncertain, according to Yeung & Yeung (2001).

Apart from the aforementioned factors, the equipment used for training include, running shoes, orthotics and running surfaces has impact on the RRI. The impact force of 2-5 times the body weight during heel-toe running was once assumed to be the main cause of running injuries (Nigg et al., 2017). To improve its shock absorption, running shoes were introduced. However, shoe technologies for additional motion-control and cushioning systems have not improved RRI (Nigg et al., 2017). There have been no studies that show that wearing running shoes reduces the risk of injury (Izquierdo et al., 2021). The introduction of orthotics was meant to correct the anatomical cushioning of the foot or lengthen the time for the foot impact while running (Taunton et al. 2002). A study conducted by Van Der Worp et al., (2015) indicated a substantial link between a history of past injury and the usage of orthotics and an increased risk of running injuries. The current results of orthotics to be a risk factor for running-related injuries should be interpreted with caution due to the lack of studies presented on relation of orthotics to running injuries (Van Der Worp et al., 2015; Nigg et al., 2017). Furthermore, a number of authors have looked into whether the running surface is a risk factor for running injuries (Taunton et al., 2003). Taunton et al., (2003) also said that when running on an unfamiliar training surface, runners

are more likely to develop an injury. There was no link between running-related injury and running on a firmer surface, according to Taunton et al. (2003) and van Gent et al. (2007). According to Lorimer and Hume (2014), increasing leg stiffness can lead to increased pre- and post-ground contact muscular activity as well as decreased joint motion. The Achilles tendon is put under more strain when jogging on softer surfaces, increasing the risk of overload and microdamage. Running on softer surfaces, such as synthetic track, sand, and grass, should be limited for athletes coming from injury (Lorimer & Hume, 2014). According to previous research, there appears to be very little data supporting the effects of running shoes and running surfaces on overuse injuries. The benefits of utilizing orthotics to prevent running-related injuries are also unknown, and the evidence is inconsistent. As a result, more research on the usage of orthotics is required.

2.6 Challenges faced by previous RRI researchers

Previous RRI researches encountered several challenges from the respect of case study data, study design and relationship between attributes. Many researchers conducted experiments and collected online self-report survey data from volunteers to conduct their research on RRI. The usage of RRI data were merely based on online self-report survey has its own set of issues, because there is a presence of recall and volunteer biasness. Kluitenberg, (2016) had mentioned that the reported injury incidence is most of the time on self-perception. Therefore, it is uncertain if RRI is analyzed from same injury interpretation.

Another significant issue that earlier researchers encountered was the difficulty in collecting huge datasets from a prospective cohort. For instance, Videbæk et al., (2015) performed meta-analysis but found only thirteen studies on running-related injury that were published between 1987 and 2014. Experimental study was bounded by ethical management where the collected data can only be studied provided with the consent of the study subject.

Many study designs used in RRI include factors leading to RRI. However, Chen et al., (2020), stated that with only a single-case design used, it weakens a research in general. Because of the highly intricate joint contact and loading conditions, batch processing of

patient-specific models is still difficult for the foot-ankle complex (Chen, Wong, Peng, et al., 2020).

Francis's (2019) study found difficulties previously identified in systematic reviews, such as a lack of consistency in designating a runner, injury, and exposure. In the areas of total number of runners, injured runners, number of injuries, and number of new injuries against recurrent injuries, there were concerns with a lack of clarity and consistency among studies.

2.7 Summary

Throughout the decades, many methods and techniques have been developed and studied to determine the body segments associating with the RRI. Many researchers confirmed that the most frequent injury locations are at the lower extremity such as the foot, ankle, knee and hip. Besides, other researches works were on the running pattern of runners.

Most research concluded with stating evidence of risk factors for RRIs is limited. Running injuries seem to have multifactorial origins. There is a need for additional high-quality studies on risk factors before strong conclusions can be drawn about the relevance of risk factors (van Poppel et al., 2020).

One common aspect shared by several related studies was identifying the potential risk factor of RRI. The results of past researchers were towards effectively minimalizing the risks associated with running.

Data analysis related to running motions are usually collected by means of self-reporting surveys, conducting experiment or by referring to publicly available data sheet with information on previous injuries, BMI, age, gender, training type, foot strike pattern, and occupation. The challenges observed in RRI were the lack of real patient data, ethical management issues and lack of clarity and consistency of past researches on injured runners.

Chapter 3 METHODOLOGY

3.1 Overview

This chapter discusses the approaches involved to prepare, pre-process and classify the dataset on running-related injury. The dataset on RRI was obtained from Kaggle, an online open dataset platform. Data mining techniques was adopted to gain knowledge from the RRI dataset with the aid of WEKA toolkit. The major task was to get the dataset ready for pre-processing and later to be applied on classification algorithms to classify the dataset into injury and non-injury classes based on the characteristics measured on the study data attributes.

3.2 Study Approach

The research was implemented across various stages as shown in the flowchart in Figure 3.1.

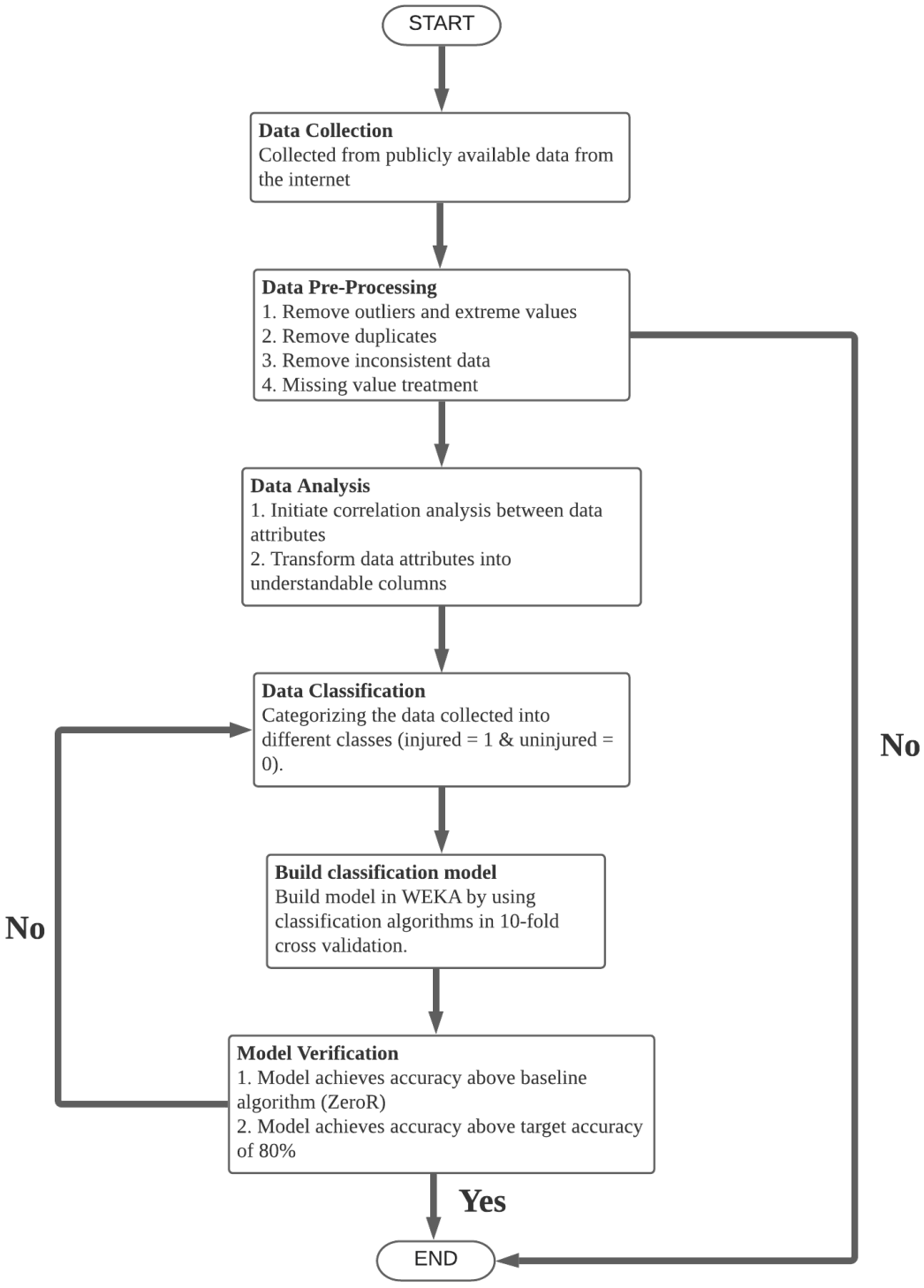

Figure 3.1 Study flow chart

3.3 Data Collection

The case study dataset was obtained from Lovdal et al., (2020). Few assumptions were made prior to using the collected dataset. First, the data set consists of random samples to ensure that each data point in the population has an equal chance of being included in the sample. Random samples are more likely to be representative of the population. So, the statistical inferences with a random sample were used. Secondly, the dataset is statistically independent, meaning that value of one observation does not influence or affect the value of other observations.

The study data consists of 74 runners (27 females and 47 males) from a training record of Dutch professional runners over a period of seven years (2012-2019). It included the middle-and-long distance runners, specifically those runners of 800 meters and the marathon. At the moment of data collection, they had been in the team for an average of 3.7 years. Most athletes competed on a national level, and some also on an international level. The study was conducted according to the requirements of the Declaration of Helsinki, and was approved by the ethics committee of the second author's institution. The experimental design decision was motivated by the fact that these groups have strong endurance-based components in their training, making their training regimes comparable.

The raw data consist of 72 attributes and 42798 instances. Table 3.1 lists the data attributes and description. Getting an understanding of the attribute names used in the dataset was essential for data pre-processing stage upon manual screening for irrelevant data attributes.

Table 3.1 Study attributes

Var	Name	Content Description
nr. sessions	**Number of sessions**	Number of trainings sessions completed.
nr. rest days	**Number of rest days**	Number of days without a training.
total kms	**Number of mileage**	Total running mileage.
max km one day	**Max km run in one day**	The maximum number of kilometers completed by running on a single day.
total km Z3-Z4-Z5-T1-T2	**Total km done in Zone 3**	The total number of kilometers done in Z3 or faster, corresponding to running above the aerobic threshold.
nr. tough sessions	**Tough session in Zone 5**	Effort in Z5, T1, T2, corresponding to running above the anaerobic threshold and/or intensive track intervals.
nr. days with interval session	**Number of days in Zone 3**	Number of days that contained a session in Z3 or faster.
total km Z3-4	**Number of km in Zone 3 and 4**	Number of kilometers covered in Z3-4, between the aerobic and anaerobic threshold.
max km Z3-4 one day	**Max km ran in Zone 3 and 4 in one day**	Furthest distance ran in Z3-4 on a single day.
total km Z5-T1-T2	**Total km ran in Zone 5**	Total distance ran in Z5-T1-T2.
max km Z5-T1-T2 one day	**Max km ran in Zone 5**	Furthest distance ran in Z5-T1-T2 on a single day.
total hours alternative training	**Time spent alternative training**	Total time spent on cross training.
nr. strength trainings	**Number of strength training**	Number of strength trainings completed.
avg exertion	**Average perceived rating**	The average rating in exertion based on the athlete's own perception of how tough each training has been.
min exertion	**Min perceived rating**	The smallest rating in exertion of all trainings of the week.
max exertion	**Max perceived rating**	The highest rating in exertion of all trainings of the week.
avg training success	**Average perceived training rating**	The average rating in how well each training went, according to the athlete's own perception.
min training success	**Min perceived training rating**	The smallest rating in training success of the week.
max training success	**Max perceived training rating**	The highest rating in training success of the week.
avg recovery	**Average perceived recovery rating**	The average rating in how well rested the athlete felt before each session.
min recovery	**Min perceived recovery rating**	The smallest rating in how well rested the athlete felt before a session.
max recovery	**Max perceived recovery rating**	The highest rating in how well rested the athlete felt before a session.

3.4 Data Preprocessing

Data pre-processing approach is to ensure only qualitative data being used for further data classification analysis. Furthermore, it is also to improve data reliability, accuracy as well as reduce data complication. Prior to using the raw dataset, it was first and foremost transformed into comma separated value (.csv) format readable by WEKA. Real-world data, is usually inconsistent, lacking or incomplete with flaws in certain ways, plus, its likely to contain unintended errors.

Firstly, the dataset attributes were analyzed via manual inspection by referring to the attached dataset description. Irrelevant data attributes (Athlete ID, Date) were removed. Data attribute (injury) was identified as the injury data class of interest to this study. The injury data class contains uninjured and injured athletes, representing data values of 0 and 1 respectively.

Initially, the raw dataset used had 73 attributes, with many redundant attributes. The redundant attributes were separated into two groups, group A and group B as shown in Table 3.2. Group A redundant attributes: nr.sessions, nr.sessions.1, nr.sessions.2, nr.sessions.3, nr.sessions.4, nr.sessions.5, and nr.sessions.6, all refers to the same number of training performed for a whole week, the attributes were averaged into a single column. Similar, initiative was carried out towards similar redundant attributes for example, total km, km Z3-4, km Z5-T1-T2, km sprinting, strength training, and hours alternative. However, for redundant attributes in group B, the data values were extracted from the original data in group A, therefore it is irrelevant for further analysis.

Table 3.2 Redundant data attributes found in dataset

A (Get average)	B (remove completely)
nr. rest days (0-6)	nr. tough sessions (effort in Z5, T1 or T2)
total km Z3-Z4-Z5-T1-T2 (0-6)	nr. days with interval session
total km Z3-4 (0-6)	max km one day

total km Z5-T1-T2 (0-6)	max km Z3-4 one day
total hours alternative training (0-6)	max km Z5-T1-T2 one day
nr. strength trainings (0-6)	min exertion
avg exertion (0-6)	max exertion
avg training success (0-6)	min training success
avg recovery (0-6)	max training success
	min recovery
	max recovery
	nr.sessions
	total kms

There also exists outliers and extreme values in the study data. Therefore, at this phase such values and data attributes will be filtered. The interquartile range filter was adopted to eliminate data that lie beyond the interquartile range as shown in Equations (3.1) and (3.2).

$$Interquartile\ Range\ (IQR) = Q_3 - Q_1$$
$$where, \quad Q_1 = lower\ quartile$$
$$Q_3 = upper\ quartile \tag{3.1}$$

$$Q_3 + 1.5(\text{IQR}) < \text{Outlier} < Q_1 - 1.5(\text{IQR}) \tag{3.2}$$

This is followed by the removal of those outliers and extreme values using RemoveWithValues filter in WEKA. This is because outliers and extreme values could possibly be an incorrect input by the dataset collector. More frequent than not, outliers and extreme values are considered low quality data that will affect both the classification results and assumptions made.

Furthermore, the attributes were examined for missing values. The dataset contained no missing value. However, the data attributes like perceived exertion, perceived

trainingSuccess and perceived recovery contained -0.01 which represents the rest days. Therefore, the training associated with the rest days was not included and discarded from further analyses.

3.5 Data Classification

Classification is the process of organizing data into labelled classes based on data records. Classifier algorithms used were BayesNet, HoeffdingTree, J48, RandomForest, RandomTree, REPTree, Naïve Bayes, IBk (k = 1, 2, 3, 4) of WEKA toolkit with ZeroR as baseline algorithm. These classifiers are chosen for its capability to predict class labels or values for the decision-making process.

Each classifier chosen was executed on three different test options, the training set, the 10-fold cross-validation, and 66% percentage split. For the training set test option, the data is tested using the training data itself. While for 10-fold cross validation, the training data is divided into 10 subsets. So, there will be 10 tests for 10-fold cross validation, where each datum will become test data once, and become training data 9 times. The classification accuracies will be averaged. For the 66% percentage split option, the classification results will be tested using 66% of the data as shown in Figure 3.3.

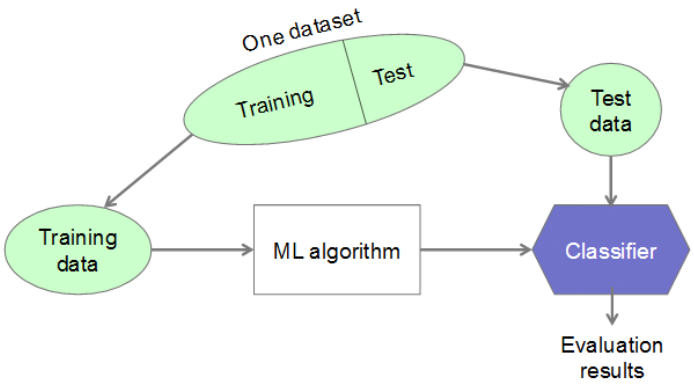

Figure 3.2 WEKA training & test data split (Martínez-Gramage et al., 2020)

3.5.1 Classification performance

The correctly performance executed in WEKA were analyzed on several indicators: Accuracy of correctly classifier instances and incorrectly classifier instances. Kappa

statistic, precision, recall, F-measure (F-score), MCC value, ROC area, as well as TPR, TNR, FPR, and FNR were observed from the confusion matrix.

Kappa statistics:

The Kappa statistics is a metric that compares an observed accuracy with an expected accuracy. The kappa statistics is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. Following, McHugh, (2012), Cohen's kappa was suggested to be interpreted as Table 3.3:

Table 3.3 Kappa statistics scale

Kappa statistics value	Agreement Interpretation
<= 0	none
0.01 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 0.99	almost perfect
1	perfect

Precision:

The precision value is the ratio of correctly predicted positive observations to the total predicted positive observations (Equation (3.3)). The closer the precision value of model to 1, the higher is the precision of model.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{(TP+FP)}} \tag{3.3}$$

Recall:

The recall value indicates the sensitivity of the model. In an imbalanced classification problem with two classes or binary classification, recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

The result is a value between 0.0 for no recall and 1.0 for full or perfect recall. The recall is calculated based on Equation (3.4).

$$\textbf{Recall (Sensitivity)} = \frac{\textbf{TP}}{\textbf{(TP+FN)}} \tag{3.4}$$

F-measure:

F-score or F-measure is a value between 0.0 for the worst F-score and 1.0 for a perfect F-score. The intuition for F-score is that both measures are balanced in importance and that only a good precision and good recall together result in a good F-score (Brownlee, 2020). F-score is calculated based on formula stated in Equation (3.5).

$$\textbf{F} - \textbf{score} = \frac{\textbf{(2*Precision*Recall)}}{\textbf{(Precision+Recall)}} \tag{3.5}$$

MCC:

Matthews Correlation Coefficient (MCC) is unlike precision, recall, and f-score, in which it takes all the cells of the confusion matrix into consideration in its formula. MCC is similar to correlation coefficient, in which is ranges between values -1 to +1. In all cases a model with a score of +1 is a perfect model and -1 is a poor model. This property is one of the key usefulness of MCC as it leads to easy interpretability. MCC is calculated based on Equation (3.6).

$$\textbf{MCC} = \frac{\textbf{TP*TN}-\textbf{FP*FN}}{\sqrt{\textbf{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}} \tag{3.6}$$

ROC area:

Receiver Operating Characteristic (ROC) area is the area under the ROC curve, abbreviated as AUC or area under curve. It is a single scalar value that measures the overall performance of a binary classifier (Hanley and McNeil, 1982). The AUC value lies between 0.5 and 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier. The scale that represents the ROC area is as stated in Table 3.4.

Table 3.4 ROC representation scale

ROC value	Represent	Grade
0.9 - 1	Excellent	A
0.8 - 0.89	Good	B
0.7 - 0.79	Fair	C
0.6 - 0.69	Poor	D
0.5 - 0.59	Fail	F

Confusion matrix:

The confusion matrix is WEKA displays four results called true positive (TP), true negative (TN), false positive (FP), and false negative (FN) in the form as shown in Figure 3.4. These results are calculated by Equations (3.7) to (3.10). In the data, the target class attribute was binary, either “not injured” or “injured”.

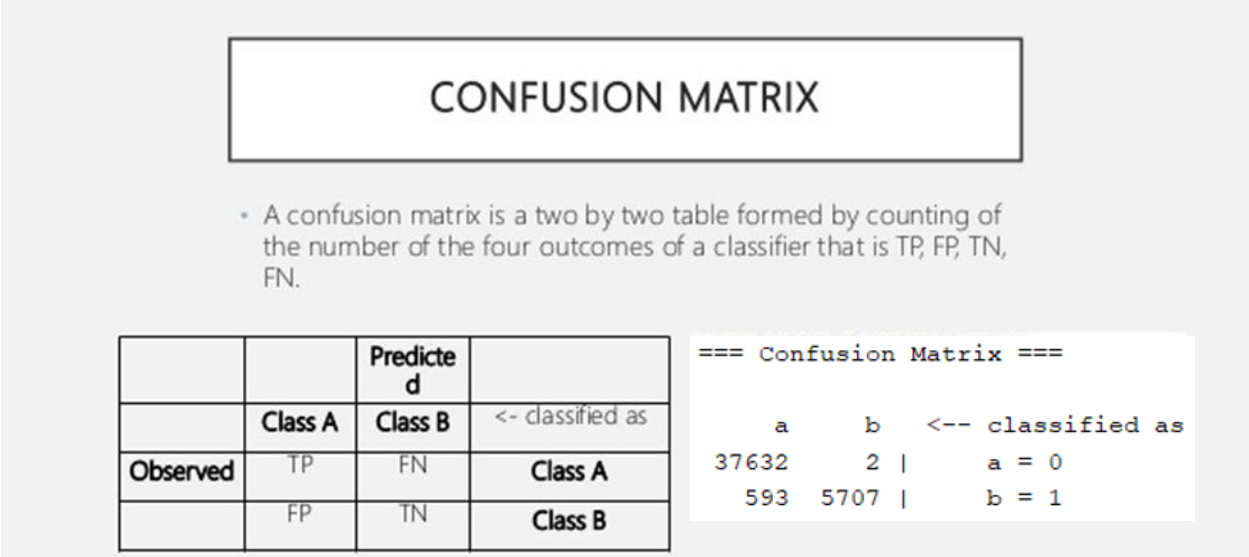

Figure 3.3 Confusion matrix form in WEKA