



Predicting running injury using kinematic and kinetic parameters generated by an optical motion capture system

Sinan Onal¹ · Mitchell Leefers¹ · Bryan Smith² · Sohyung Cho¹

© Springer Nature Switzerland AG 2019

Abstract

Although runners are at high risk of back and lower extremity injuries, available tools detect only current injury. Here, a model was developed to analyze kinetic and kinematic running gait data collected by an optical motion capture system to predict future injuries based on an individual's running gait pattern. The two key points, when the joints are most vulnerable because internal forces are the greatest, in the continuous running gait cycle were used to extract average parameter values to create predictive models: the heel strike, and when one leg supports the body weight. Three different prediction models—logistic regression, random forest, and boosting—were built using 10 significant parameters identified in a two-step feature selection approach. All collected metric data were normalized before building the predictive models to avoid outlier values and redundancy. The three models were tested to determine whether they could predict that a participant would incur chronic running injuries in the future based on their current running gait pattern. The logistic regression model had the highest prediction accuracy: the area under the curve was 0.9016 [95% confidence interval (CI) 0.8808–0.9369] for logistic regression, 0.8892 (95% CI 0.8463–0.9152) for the random forest, and 0.8732 (95% CI 0.8401–0.9178) for boosting. Further model development may not only enable clinicians to integrate injury intervention into running programs but also lead to predictive models that recognize patterns associated with neurological disorders, such as Parkinson's disease, autism, and multiple sclerosis, in which gait and balance deficiencies may be symptoms or even predictors of disease.

Keywords Kinematics · Running injury · Prediction · Optical motion capture

1 Introduction

A large number of runners at every level of expertise incur injuries that may be linked to several different factors, such as age, body mass index, distance, experience, previous running injuries, incomplete healing, and faulty biomechanics. Of the different types of injuries among runners, back and lower extremity injury rates are high [1, 2]. Pre-participation screening tools, such as the Functional Movement Screen and Star Excursion Balance Test, can be used to identify current injury in many runners [3]. However, rather than screening only for current injuries, a gait screening tool that could predict whether an individual

may incur a chronic running injury in the future would substantially reduce the financial, health, and psychological burden.

Analyses of human gait have been performed by health professionals using numerous techniques in diverse fields of healthcare for various reasons. In sports medicine, gait analysis is often performed on athletes to optimize their movements for energy use and to avoid dangerous patterns that may cause injuries both while conditioning and performing. Advances in motion capture and analysis technology have allowed for extremely accurate computerized 3-dimensional (3-D) human body models with six degrees of freedom to be built [4]. These models enable

✉ Sinan Onal, sonal@siue.edu | ¹Department of Mechanical and Industrial Engineering, Southern Illinois University Edwardsville, Edwardsville, IL, USA. ²Department of Applied Health, Southern Illinois University Edwardsville, Edwardsville, IL, USA.



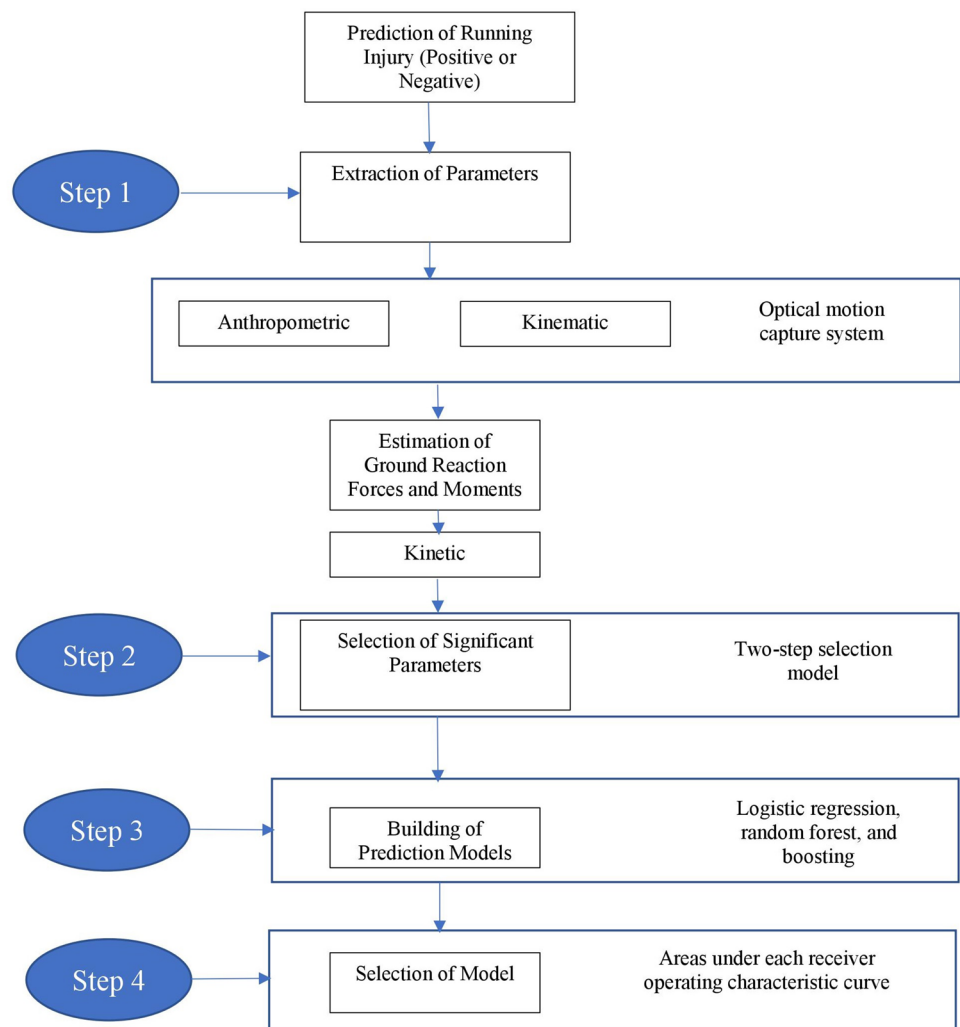
users to extract kinematic and kinetic data for analysis of the human gait. Although these types of data are useful for healthcare professionals observing an individual's gait, the analysis of the data is still left in the hands of the healthcare provider. This type of analysis takes time, is prone to errors, and can be subjective because there is no set standard for a "normal" gait. Thus, a completely objective method that can predict whether an individual would incur a chronic running injury in the future based on the patterns found in kinematic and kinetic gait data collected from a motion capture system is necessary.

The aims of the present study were to determine the most important kinematic and kinetic parameters extracted during a running gait using a motion capture system and to develop a model that incorporated these parameters to predict a future running injury. The model proposed in the present study was developed using four steps: (1) extract the parameters (2) select the significant

parameters using a two-step model, (3) build three prediction models (logistic regression, random forest, and boosting) to predict whether the individual would incur chronic running injuries in the future based on their current running gait patterns, and (4) select the appropriate model (Fig. 1). The results of the present study will enable clinicians to integrate suggested interventions into an individual's running program to prevent immediate injury and decrease the probability of future running injuries.

Section 2 of this article describes the methods used to ascertain the model, including participant recruitment, data collection, and data analysis. Section 3 provides the results and evaluates the performance of the model, and Sect. 4 offers a discussion of the model, including the strengths and limitations of the present study and model and future directions.

Fig. 1 Steps used to develop the proposed model



2 Methods

2.1 Definition of injury

Various definitions of injury have been used throughout the literature. In the present study, a running-related injury in recreational runners was defined as stated by Yamato et al.: “running-related (training or competition) musculoskeletal pain in the lower limbs that causes a restriction on or stoppage of running (distance, speed, duration, or training) for at least 7 days or 3 consecutive scheduled training sessions, or that requires the runner to consult a physician or other health professional” [5].

2.2 Instruments and procedures

The present study was approved by the Southern Illinois University, Edwardsville (SIUE) Institutional Review Board. Participants received verbal explanations about the purpose of the study and the methods before providing informed consent. All participants then signed informed consent forms.

The data for this research was collected in the Motion Capture and Analysis Laboratory on the SIUE campus. The laboratory houses a Vicon optical motion capture system (Oxford Metrics, Oxford, UK) and all the necessary software tools to extract kinematic and kinetic data from motion capture trials. The Vicon Vantage V5 standard camera is a 5-megapixel camera that captured images at 420 frames/s. The Vicon Vue video camera enabled the capture of a reference video. The Vicon Nexus 2.6 software package (Oxford Metrics, Oxford, UK) was used for system calibration, data capture, post-processing and analysis, and data export. Tracker 3.0 software (Oxford Metrics, Oxford, UK) allowed the integration of the data into a 3-D visualization application, Visual 3-D V5 Professional software (C-Motion, Germantown, MD, USA).

Collecting the data for each participant was conducted in a systematic manner to ensure the most accurate representation for each person’s running gait cycle. First, each participant was asked to provide a static trial in which he or she stood in the middle of the laboratory with the arms in front of the body for a brief period. This trial was used to ensure that all the markers were in the correct locations and were visible to the cameras. This trial was also used later while analyzing the data to build a model based on specific anthropometric measurements, including mass, height, leg length, knee width, ankle width, shoulder offset, elbow width, wrist width, and hand thickness. After the static trial, each participant provided five dynamic trials. These trials consisted of the

participant jogging on a treadmill (ProForm Performance 300i) for 45 s. Running velocity affects lower extremity kinematics [6]. Therefore, matching treadmill speed to a similar speed at which an injured runner experienced symptoms was accommodated. For an injury-free runner, the treadmill speed was set to match the running velocity of a “long run,” because if this group of runners demonstrated abnormal biomechanics during the longer runs, the faults would accumulate over the longer exercise period and might contribute to running injuries [7]. After these dynamic trials, Nexus, version 2.6, software was used to rebuild the trial, that is, the software recreated the set of markers on the monitor and played back the entire trial. The markers were then labeled with respect to their specific location.

2.3 Parameter extraction and significant parameter selection

After the continuous data were computed, a method in the software called “event creation” was used. This method allowed a user to create events in a trial based on specific parameters within the trial. Because force plates were not used in this study, the gait events were created manually by using a method commonly referred to as a coordinate-based algorithm [8]. The first step in this process was to transform the markers used for the heel and the toe on each foot into the coordinate system of the pelvis to create a parameter that computed the distance from the pelvis to the heel and toe. Once those parameters were computed, they were used to define a maximum and minimum distance of the markers with respect to the pelvis segment. Thus, the heel strike event of the gait cycle was defined as the maximum distance the heel marker translated forward (anteroposterior), and the toe-off event of the gait cycle was defined as the minimum distance of translation (anteroposterior) for the toe marker. Figure 2 depicts the coordinate system used in this study. The three key phases of running have been previously described [7] as (1) the end of the terminal swing, identified as when the foot remains elevated from the treadmill, just before initial contact, (2) initial contact, identified as when the foot hits the ground, and (3) the loading response,

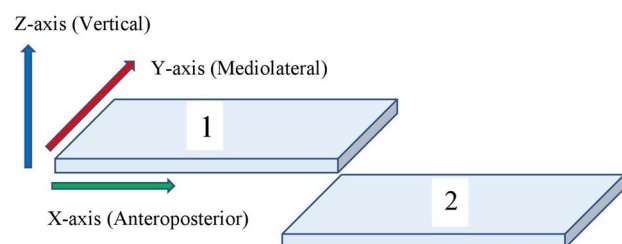


Fig. 2 Coordinate system

identified as when the runner's weight is being transferred onto the lead leg and is characterized by the presence of shoe deformation.

After the events of the gait cycle were determined, it was then possible to extract data values at or between gait events. Kinematic patterns are very similar for treadmill running and over-ground running [9]. All the necessary kinematic gait parameters, that is, the absolute and relative angles at these events, were extracted from the pelvis, hip, knee, and ankle bilaterally based on the position of the reflective markers. The parameters were then transformed into a discrete form. In addition, because there were many gait cycles in each trial (36 ± 4.3 cycles), the means and standard deviations of those parameters were computed from each trial for the left and right sides of the pelvis, hip, knee, and ankle.

Assessment of ground reaction forces and moments is an important stage in the biomechanical analysis procedure [10]. Conventionally, these measures are recorded during running using a treadmill that has force plates as an integral part of it, which is very costly. As an alternative, kinetic parameters during the running gait can be estimated with inverse dynamics if anthropometric and kinematic data are known. Because a treadmill with integrated force plates was not used in the present study, ground reaction forces and moments data were estimated using a previously proposed accurate method [11]. To predict ground reaction forces without a force plate, the traditional method of Newtonian mechanics was used for the single support phase. An artificial neural network model was applied for the double support phase to solve statically indeterminate structure problems.

The proposed model computed kinetic and kinematic metrics using six degrees of freedom, thus allowing the degree of rotation about every axis to be computed on the pelvis, hip, knee, and ankle. However, not all the metrics concerning these axes might be of importance for the prediction model. Therefore, after kinematic and kinetic parameter extraction, a two-step significant parameters subset selection was performed. In the first step, irrelevant or redundant features were removed using the independent significance parameter selection method as previously described [12]. This step eliminated parameters with a significance level lower than 2 as calculated using the following equation:

$$significance_i = \frac{|mean(Injured_i) - mean(Normal_i)|}{\sqrt{\frac{var(Injured_i) + var(Normal_i)}{n_1 + n_2}}}, \quad (1)$$

where $Injured_i$ indicates the i th parameter being measured from the injury data set, $Normal_i$ represents the i th

parameter being measured from the normal data set, and n_1 and n_2 are the corresponding number of injured and uninjured participants, respectively. In the second step, a number of subset parameters were selected for the prediction models using the sequential forward selection method measured by fivefold cross-validation. In this fivefold cross-validation, the parameter set was divided into five subsets of equal size. Each subset was tested on the remaining four subsets using the mean squared error that minimized the mean criterion value. This process continued until the addition of one more parameter did not decrease the criterion any further. After this step, the 10 most significant parameters were identified to build the prediction models.

2.4 Building prediction models

The first predictive model built was a logistic regression model [13] using the selected significant parameters. Bootstrapping methods, using 1000 iterations, were used to resample the full parameters set to reduce bias incurred by uneven numbers of the data set originating from each participant, and to enable model development and validation using the same data set. For each iteration, mean areas under each receiver operating characteristic (ROC) curve (A_{UC}) were calculated using the selected significant parameters and models of interest; mean A_i and bootstrap 95% confidence intervals (95% CIs) were reported; differences between A_{UC} s were calculated, and the bootstrap p value determined.

Two machine learning methods were also used, random forest [14] and boosting [15], to build the prediction models. Then, all three prediction models were compared, and the best predictive model was selected as a prediction model for running injuries.

A random forest is a tree-based learning model. The basic premise of the random forest is that a number of decision trees are built, and each individual tree is used to make a prediction. This process starts with bootstrapping. Bootstrapping is the process of taking multiple samples from a single set of training data. After bootstrapping, a method called bootstrap aggregation (bagging) is used to lower the variance within each individual tree that is built. Bagging works on the principle of averaging a set of observations to reduce variance. The prediction can be calculated as $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^R(x)$, with R number of training sets, and the predictions averaged using $\hat{f}_{avg}(x) = \frac{1}{R} \sum_{r=1}^R \hat{f}^r(x)$. Each tree is then modelled on the r th bootstrapped training set to obtain $\hat{f}^{*r}(x)$. The average of the predictions is obtained using $\hat{f}_{bag}(x) = \frac{1}{R} \sum_{r=1}^R \hat{f}^{*r}(x)$. The random forest model is able to decorrelate the trees by considering only a subset of predictors in the data for each split. In

other words, not every predicting feature is considered in each split. This ensures that a new sample of predictors is chosen at each split, thus decorrelating the predictions by ensuring that each tree is constructed differently. The number of m predictors considered for each split in this model was $m = \sqrt{p}$, where p is the total number of predictors in the data set. After the training process was completed for the model, the final prediction was made on a majority vote basis. That is, each decision tree built made a prediction, and the most frequent prediction that was made was the final prediction of the model.

Boosting is an iterative process that focuses on misclassified data such that each tree is based on the weighted average of the data points, and the weights are calculated based on the previous model in the iterative process. The random forest and boosting methods were validated in the present study using an approach with tenfold cross-validation and 100 times replication.

3 Results

In total, 36 individuals participated in the present study: 14 participants (age, 21.2 ± 3.4 years; body mass, 71.9 ± 3.8 kg; and height, 185.4 ± 11.4 cm) had a previous injury or were currently experiencing a chronic injury from running, such as shin splints, knee pain, or lower back pain, and 22 participants (age, 23.6 ± 4.1 years; body mass, 75.9 ± 4.6 kg; and height, 192.7 ± 21.0 cm) were injury-free at the time of data collection. All participants were active recreational runners who completed a minimum of three training sessions no less than 30 min per week. Each participant provided five trials of running gait data. This sample size provided more than 80% statistical power.

Tables 1 and 2 give the kinematic and estimated kinetic parameters extracted from injury-free and injured participants.

Figure 3 shows the relative and absolute angles of the pelvis, hip, knee, and ankle. The x-axis of all panels in Fig. 3 represents the gait cycle, which is the time frame between two consecutive heel strikes. A cycle consists of two main phases: the stance (support) and swing phases. The stance phase starts with the heel strike and ends with the toe-off. The swing phase starts with the toe-off and ends with the heel strike. This means that on the x-axis, from 0 to 50 represents the stance phase of the gait cycle, and from 50 to 100 represents the swing phase. Each colored line on the graph represents the data for a cycle, and the thicker black line represents the mean values for all the cycles from the trials.

The two-step significant parameters selection method was implemented after this step. These significant parameters were used to build the prediction models. The significant features were angular velocities and moments along the y-axis and the z-axis. These parameters are important because they refer to the abduction or adduction (anteroposterior) of the joints and the internal rotation (vertical) of the joints and the rate at which these rotations are occurring during the gait cycle. Angular velocity was calculated as the derivative of the joint angles with respect to time. Another important parameter that was used in the analysis was the internal reaction forces at the knee and ankle joints. These forces were computed using a 3-D force vector. To avoid focusing on the reaction along one axis, the resultant force was computed and used in the prediction model. These parameters were first computed as continuous values for the entire gait cycle, and the data were normalized. Box 1 shows the most significant parameters used to build the prediction model.

After selecting the best descriptive set of parameters, the predictive models were built using a logistic regression model, a random forest, and boosting. Mean areas under each ROC curve (A_{UC}) were calculated using bootstrapping techniques for data resampling and were used during model development and performance comparisons. The A_{UC} -ROC curve is a performance measurement for classification problems at various thresholds settings. The ROC is a probability curve, and A_{UC} represents the degree or measure of separability, providing how much each model is capable of distinguishing between classes. The higher the A_{UC} , the better the model is at predicting zeros as zeros and ones as ones. By analogy, the higher the A_{UC} , the better the model is at distinguishing between participants with injury and those with no injury.

The training algorithm used to build the random forest model was the split, train, and test method from the scikit-learn utility package used in the Python, version 3.7, programming language. This algorithm allows for the data to be randomly split into training and testing data (70% was used as training data, 30% used as testing data). After the model was trained and tested, the accuracy score, confusion matrix, and variable importance were all calculated to further understand the model that was built.

Table 3 and Fig. 4 show the A_{UC} -ROC results for the three prediction models built using different parameter sets, baseline data, parameters after the first parameter selection step, and parameters after the second parameter selection step to differentiate participants by their condition.

Table 1 Kinematic parameters

Kinematic parameter	Phase	Injury-free population				Injured population			
		Mean		SD		Mean		SD	
Pelvis, degrees									
Maximum pelvic tilt	Cycle	3.7		0.8		18.3		1.2	
Minimum pelvic tilt	Cycle	−1.6		1.2		11.8		0.8	
Maximum pelvic obliquity	Cycle	3.9		1.6		3.5		0.6	
Minimum pelvic obliquity	Cycle	−6.9		1.4		−2.9		0.5	
Maximum pelvic rotation	Cycle	4.7		1.1		4.1		0.7	
Minimum pelvic rotation	Cycle	−1.3		2.1		−4.3		1.0	
		Left		Right		Left		Right	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Hip, degrees									
Maximum hip flexion	Stance	15.1	3.2	17.2	2.1	36.7	0.3	37.2	1.0
Maximum hip flexion	Swing	24.3	2.1	27.9	1.5	37.5	2.8	39.3	1.9
Maximum hip extension	Stance	−11.3	1.6	−16.2	1.0	−2.9	1.1	−7.2	2.2
Sagittal hip range of motion	Cycle	36.2	1.4	40.3	4.8	41.6	2.5	46.8	2.0
Maximum hip adduction	Stance	1.1	0.9	8.8	1.4	3.3	0.4	1.5	0.4
Maximum hip abduction	Stance	−9.3	1.3	−4.9	1.3	−3.6	0.3	−6.5	1.4
		Left		Right		Left		Right	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Knee, degrees									
Maximum knee flexion	Stance	54.3	0.7	47.9	1.2	45.8	0.8	40.9	4.4
Maximum knee flexion	Swing	63.8	2.0	69.3	0.7	67.2	1.0	67.1	1.2
Maximum knee extension	Stance	3.0	0.4	7.8	1.4	0.4	0.4	3.2	1.4
Sagittal knee range of motion	Cycle	65.1	2.9	64.6	3.3	67.3	1.6	63.1	2.1
		Left		Right		Left		Right	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Ankle, degrees									
Maximum ankle dorsiflexion	Stance	3.2	1.9	1.5	0.9	−5.0	1.1	−4.4	1.3
Maximum ankle dorsiflexion	Swing	−6.0	0.5	−7.2	1.1	−9.4	1.1	−8.5	1.7
Maximum ankle plantar flexion	Stance	−18.4	1.7	−32.0	2.1	−27.7	2.0	−27.3	2.0
Sagittal ankle range of motion	Cycle	25.0	5.2	33.4	5.8	26.2	2.0	36.0	1.6
Maximum foot rotation	Cycle	0.4	1.2	7.6	2.7	3.5	0.7	3.9	1.6

Without parameter selection steps, the A_{UC} was 0.6339 (95% CI 0.5905–0.6598 using logistic regression, 0.6428 (95% CI 0.6212–0.6646) using a random forest, and 0.7168 (95% CI 0.7001–0.7443) using boosting (Fig. 4, left panel). The boosting-based classification model was able to predict better than the other two models. The random forest was the second-best predictor. However, the prediction percentages were very low. After the first parameter selection step, the A_{UC} was

0.7440 (95% CI 0.7254–0.7786) using logistic regression, 0.7555 (95% CI 0.7412–0.8093) using random forest, and 0.8677 (95% CI 0.8087–0.9104) using boosting (Fig. 4, middle panel). The boosting-based classification model was again able to predict better than the other two models. The random forest was again the second-best predictor. However, the prediction percentages were still very low. After the second parameter selection step, the prediction results were markedly improved, with the

Table 2 Estimated kinetic parameters

Kinetic parameter	Injury-free population				Injured population			
	Left		Right		Left		Right	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Anteroposterior GRF (N/kg)	6.98	3.53	5.99	2.90	7.27	2.31	2.70	3.02
Mediolateral GRF (N/kg)	4.78	4.23	4.75	1.47	2.06	2.66	3.95	2.51
Vertical GRF (N/kg)	16.62	3.22	5.27	1.62	14.53	0.30	8.33	0.87
Frontal GRM (Nm/kg)	1.75	0.43	1.28	0.14	1.13	0.22	0.69	0.14
Sagittal GRM (Nm/kg)	3.60	1.50	3.01	1.19	3.39	2.08	2.04	1.00
Transverse GRM (Nm/kg)	1.17	0.32	0.78	0.20	0.48	0.94	0.64	0.32
Ankle flexion moment (Nm/kg)	3.31	1.15	2.05	0.95	2.08	2.68	3.23	0.17
Knee flexion moment (Nm/kg)	2.48	0.98	1.18	0.77	2.11	1.61	0.65	0.92
Hip flexion moment (Nm/kg)	2.73	0.87	1.68	0.43	1.97	1.54	0.49	0.13
Hip abduction moment (Nm/kg)	0.38	0.69	0.31	0.64	0.21	0.30	0.23	0.74
Hip extension moment (Nm/kg)	1.86	0.45	0.97	0.35	1.47	1.33	0.12	0.33
Knee varus moment (Nm/kg)	1.69	0.65	1.40	0.28	0.64	1.06	0.15	0.22
Ankle plantarflex moment (Nm/kg)	3.62	2.92	2.27	1.67	3.34	2.06	1.39	3.02
Knee resultant force (N/kg)	77.49	29.61	52.82	17.27	60.44	11.26	78.24	27.78
Ankle resultant force (N/kg)	167.54	34.88	141.17	35.87	142.71	19.76	157.26	19.39
Hip resultant force (N/kg)	95.36	19.16	90.09	14.99	84.43	11.14	99.09	14.61

GRF ground reaction force, GRM ground reaction moment

A_{UC} 0.9016 (95% CI 0.8808–0.9369) using logistic regression, 0.8892 (95% CI 0.8463–0.9152) using a random forest, and 0.8732 (95% CI 0.8401–0.9178) using boosting (Fig. 4, right panel). The logistic regression-based classification model was the best predictor, followed by the random forest and boosting models.

4 Discussion

In the present study, kinetic and kinematic parameters extracted using an optical motion capture system were combined to build a prediction model to accurately predict whether a participant was injured or not. Of the three different prediction models developed, the logistic regression model had the highest prediction accuracy. This model will be improved in future studies to create a model that accurately predicts the probability that a person will incur a chronic running injury based on their current running gait. To the best of our knowledge, there is no study that integrates a range set of kinematic and kinetic parameters to predict future running injury using an optical motion capture system.

The present study has a few limitations that should be considered when interpreting the results. Because of

the lack of having a force sensing tandem treadmill in our laboratory, kinetic parameters have been estimated using a model that might produce an error. For more accurate analyses, this type of treadmill should be integrated into the study. In addition, it was assumed that GRFs and GRMs were the only external forces applied to the individuals during the running activity. However, in a wider spectrum, there might be secondary external forces introduced in the activity. Finally, the proposed model was dependent on a set of reflective markers, which requires 21 reflective markers. In future studies, minimizing the number of markers to make the prediction model more practical for clinical applications should be investigated.

With additional data capture and analysis, more accurate and efficient prediction models can be constructed and applied. This type of predictive analysis pertains to more than just running gait analysis for sports medicine or physical therapy applications. The human gait holds much information about the neurological state of a person. The gait and the ability to balance have been studied in many cases for various neurological disorders, including Parkinson's disease, autism, and multiple sclerosis. These are examples of neurological disorders in which gait deficiencies have been defined as symptoms and in some cases have

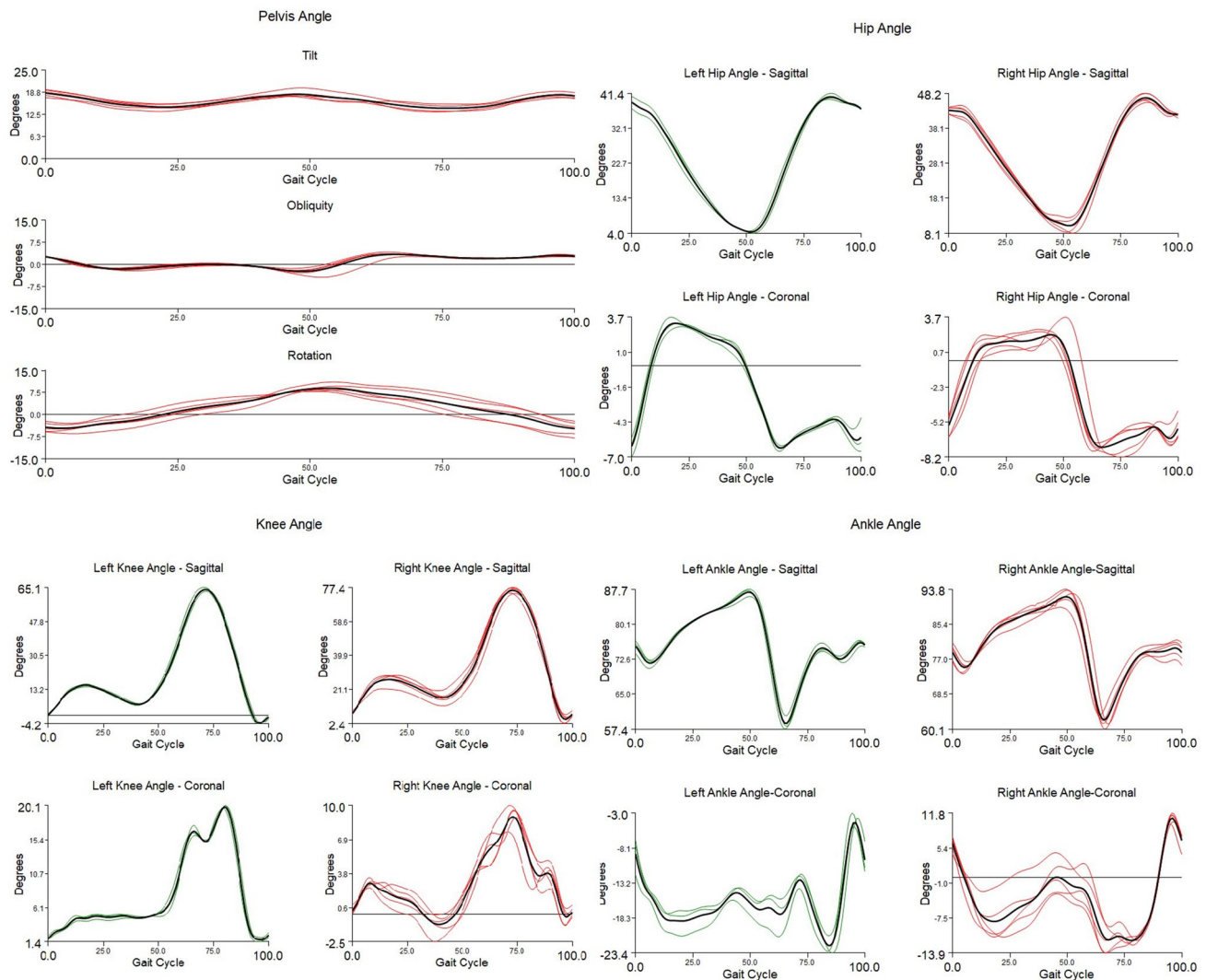


Fig. 3 Descriptive kinematic parameters from left and right sides extracted using the motion capture system

Box 1 The most significant features selected by the two-step technique

Selected significant parameters

- Right/left ankle resultant force at RHS/LHS (N)
- Right/left knee resultant force at RHS/LHS (N)
- Right/left knee extension during stance (degree)
- Right/left knee flexion during stance (degree)
- Hip extension during stance (degree)
- Pelvic tilt (degree)

RHS right heel strike, *LHS* left heel strike

been shown to be predictors. With the use of motion capture, either through a worn device or an optical motion capture, extremely accurate data is obtained. Such data hold information that is unable to be seen by

the human eye and can be used to develop predictive models with the ability to recognize patterns involved with neurological disorders.

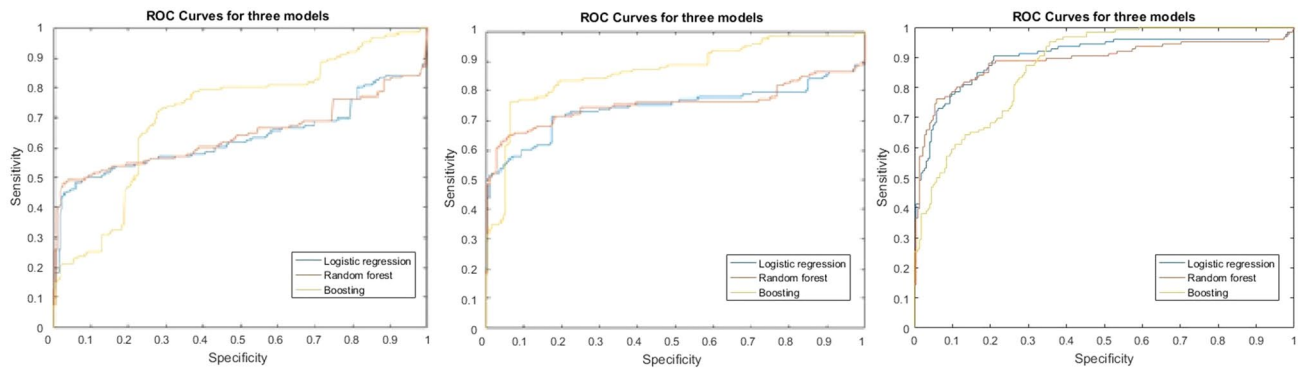
5 Conclusions

To summarize, running injuries are a common problem for runners. The purpose of the present study was to use a motion capture system to determine the most important running gait parameters and to develop a prediction model that incorporates these parameters to predict a likely future chronic running injury. The proposed model was developed using parameter extraction and two steps to select significant parameters to build three prediction models (logistic regression, random forest, and boosting).

Table 3 Accuracy comparisons of three different prediction models using different sets of parameters

Condition	Mean area under the receiver operating characteristic curve (A_{UC})		
	Logistic regression	Random forest	Boosting
Without parameter selection steps (28 parameters)	0.6339	0.6428	0.7168
After the first parameter selection step (12 parameters)	0.7440	0.7555	0.8677
After the second parameter selection step (10 parameters)	0.9016	0.8892	0.8732

Bold value indicates highest accuracy

**Fig. 4** Receiver operating characteristic (ROC) curves for logistic regression, random forest, and boosting prediction models

The proposed model showed potential for accurately predicting future injury by analyzing the running gait patterns of runners.

Acknowledgements We thank the SIUE Undergraduate Research and Creative Activities Program for their financial support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- van der Worp MP, ten Haaf DSM, van Cingel R, de Wijer A, Nijhuis-van der Sanden MWG, Staal JB (2015) Injuries in runners; a systematic review on risk factors and sex differences. *PLoS ONE* 10(2):011
- Taunton JE, Ryan MB, Clement DB, McKenzie DC, Lloyd-Smith DR, Zumbo BD (2002) A retrospective case-control analysis of 2002 running injuries. *Br J Sports Med* 36(2):95–101
- Sanders B, Blackburn TA, Boucher B (2013) Preparticipation screening - the sports physical therapy perspective. *Int J Sports Phys Ther* 8(2):180–193
- Zelik KE, Takahashi KZ, Sawicki GS (2015) Six degree-of-freedom analysis of hip, knee, ankle and foot provides updated understanding of biomechanical work during human walking. *J Exp Biol* 218(6):876–886
- Yamato TP, Saragiotto BT, Lopes AD (2015) A consensus definition of running-related injury in recreational runners: a modified Delphi approach. *J Orthop Sports Phys Ther* 45(5):375–380
- Brughelli M, Cronin J, Chaouachi A (2011) Effects of running velocity on running kinetics and kinematics. *J Strength Cond Res* 25(4):933–939
- Souza RB (2016) An evidence-based videotaped running biomechanics analysis. *Phys Med Rehabil Clin N Am* 27(1):217–236
- Zeni JA Jr, Richards JG, Higginson JS (2008) Two simple methods for determining gait events during treadmill and overground walking using kinematic data. *Gait Posture* 27(4):710–714
- Riley PO, Dicharry J, Franz J, Della Croce U, Wilder RP, Kerrigan DC (2008) A kinematics and kinetic comparison of overground and treadmill running. *Med Sci Sports Exerc* 40(6):1093–1100
- Karatsidis A, Bellusci G, Schepers HM, de Zee M, Andersen MS, Veltink PH (2016) Estimation of ground reaction forces and moments during gait using only inertial motion capture. *Sensors (Basel, Switzerland)* 17(1):75
- Oh SE, Choi A, Mun JH (2013) Prediction of ground reaction forces during gait based on kinematics and a neural network model. *J Biomech* 46(14):2372–2380
- Onal S, Lai-Yuen S, Bao P, Weitzenfeld A, Hart S (2016) Automated localization of multiple pelvic bone structures on MRI. *IEEE J Biomed Health Inform* 20(1):249–255
- Cox DR (1958) The regression analysis of binary sequences. *J R Stat Soc Ser B (Methodol)* 20(2):215–242
- Tin Kam H (1995) Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*
- Schapire RE (1990) The strength of weak learnability. *Mach Learn* 5(2):197–227

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.