Indian Journal of Engineering & Materials Sciences
Vol. 32, December 2025, pp. 727-734
DOI: 10.56042/ijems.v32i06.18481

Identification of foot typology from gait analysis using machine learning approaches

Saraswathy Gnanasundaram[a]*, Nithya Rajagopalan[b], Jeni Christina George[b], Akileshwar Ganeshan[b] & Krishnapriya Sundararaman[b]

[a]Footwear Biomechanics Unit, CSIR-Central Leather Research Institute (CSIR- CLRI), Sardar Patel Road, Adyar 600 020, Chennai, India
[b]Department of Biomedical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam 603 110, Chennai, India

Received: 03 April 2025; accepted: 23 October 2025

Human foot has unique anatomical and functional structure to facilitate various movements. It is difficult to classify the foot typology by visual assessment and by using static analysing tools. The aim of the study is to predict foot typology and classify the foot type using instrumental gait analysis and machine learning techniques, focusing on women aged 20-29 and 40-49 years to identify key features contributing to conditions like flat feet and high arch feet, and to facilitate early detection and correction of abnormal foot typologies. Data collection has involved 25 participants in Group 1 (aged 20-29 years) and 15 participants in Group 2 (aged 40-49 years), each undergoing 3 trials in instrumented treadmill gait analysis (ITGA), alongside questionnaires, consent forms, and body composition analysis. Participants have been selected based on specific inclusion and exclusion criteria to ensure valid results. The data has been used machine learning models to classify foot typology into Normal Arch, High Arch, and Flatfoot. The dataset has been separated as 70% training and 30% testing for multi-class classification. The foot typology has been classified using machine learning models based on 49 features of gait analysis. The features identify the patterns and differences among the different foot typologies for accurate classification. The application of synthetic minority over-sampling technique (SMOTE) to balance the dataset also improved the accuracy across all models. Among different machine learning models employed, bagging algorithm has achieved the highest accuracy of 95.6% and 94.3% for Group 1 and Group 2 respectively. The study has indicated effectiveness of machine learning techniques in classification of foot typology using gait analysis proving valuable insights into foot biomechanics and improved precision of diagnosis for personalized intervention strategies.

Keywords: Biomechanics, Foot typology, Gait, Instrumental treadmill analysis, Machine Learning

1 Introduction

Human movements such as standing, walking, and running is facilitated by a unique anatomical and functional structure, the Foot[1,2]. The foot plays a vital role in maintaining the postural stability (PS), efficient shock absorption and optimization of impact mitigation during these movements[3-5]. The difference in structural and functional alignment of the foot is described as a foot type (FT). The human foot type influences the functions of foot and lower limb as well as kinematic and kinetic gait parameters[6]. Foot typology refers to the classification of feet according to shape, structure, and alignment. The three major categories of foot types based on the foot arch structure are: Normal arch foot, High arch foot and Flat foot.

The arch structure is designed to bear the weight of the body and absorb impact with ground that is produced with locomotion[7,8]. The flexibility of the foot due to the presence of arches facilitates everyday loco-motor functions such as walking and sprinting[9]. Hence a properly functioning arch helps distribute forces evenly during human locomotion promoting efficient propulsion and reducing the risk of injuries[10,11]. However, variations in arch height and structure can lead to biomechanical imbalances, impacting load distribution, foot strike pattern on the ground, and force absorption during movement[12].

Various factors contribute to foot topology including anatomical reasons, genetics, and pathological factors. External factors such as footwear choices, injuries, and lifestyle habits also influence foot typology. Considering the diverse factors influencing foot typology, conditions such as flatfoot and high arch are regarded as abnormal variations. In flat feet, the arch of the foot is lowered, resulting in the high area of contact of sole with the ground, foot

*Corresponding author (E-mail: saraswathyg.clri@csri.res.in)

pronation and knee valgus causing pain, discomfort and poor posture. Genetics, diseases like diabetes, arthritis, or foot injuries, and risk factors including pregnancy, obesity, and certain medical conditions are some of the causes for an individual to develop flat feet[13]. Flat foot can result in abnormal weight distribution leading to back and leg pain, and persistent heel and arch pain that worsens with activity[14,15]. Flat feet can also cause greater muscle tension and fatigue while walking due to the uneven distribution of weight and pressure on the foot[16].

High arch feet on the other hand are characterized by abnormally elevated arch structure and low midfoot area contact with the ground. High arch feet reduce shock absorption and increase injury risk, leading to ankle instability and foot pain, all of which can affect overall mobility and quality of life[17]. Although high arches provide greater stability, it may lead to reduced pressure absorption and increased stress on the foot during activities[16,17]. Individuals with high arch feet are prone to conditions such as plantar fasciitis, metatarsalgia, stress fractures, ankle sprain, hammer toes, Achilles tendinitis and corns.

The higher prevalence of abnormal foot typologies like flat feet and high arch feet in women compared to men is influenced by several factors including anatomical differences, footwear choices, occupational demands, and hormonal changes. Anatomically, women have a broader pelvis and more hip abduction, impacting stride length and gait mechanics, along with a weaker muscle-ligament structure and smaller, more rounded metatarsal heads, which are more prone to deformities. Hormonal changes during pregnancy cause ligaments to relax and arches to flatten, contributing to foot abnormalities[18-21]. Footwear choices, particularly high-heeled and ill-fitting shoes, play a significant role in developing conditions such as hallux valgus and varus deformity of the fifth toe. The frequent changing of shoes to follow fashion trends often leads to wearing poorly constructed footwear, exacerbating foot issues. Occupational factors, including dress code requirements that mandate high-heeled shoes and prolonged standing, shift the centre of gravity forward and adversely affect body contouring[22]. Women have a genetic inclination towards development of high arch and flat feet. The combination of anatomical and genetic factors results in a higher incidence of foot typologies in women, impacting their overall foot health, and thus increasing the risk of musculoskeletal conditions such as osteoporosis and osteoarthritis[23].

This study emphasizes and highlights the importance of foot typology, an often-overlooked aspect, in maintaining lower limb health. Machine learning techniques is used to evaluate foot typology to enhance individual gait patterns. Given the unique biomechanics of women and their high susceptibility to musculoskeletal conditions, the study focuses on women's gait across age groups Group 1 (20-29 years) and Group 2 (40-49 years). Using machine learning models, the gait parameters are used in the classification of foot typology and the data driven approach facilitates for early detection of abnormal foot typology and implement corrective measures at the earliest.

2 Materials and Methods

The study's methodology, depicted in Fig. 1, adopts a systematic approach to evaluate foot typology and its impact on gait patterns, with a particular focus on women in two age groups (20–29 and 40–49). The first step of the study begins with data collection,

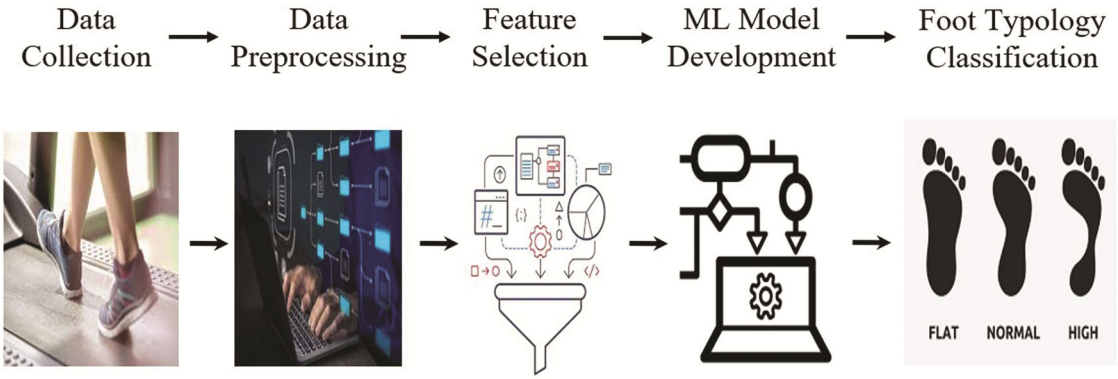

Fig. 1 — Methodology (a) data collection, (b) data preprocessing, (c) feature selection, (d) ml model development, and (e) foot typology classification.

followed by data preprocessing methods to organise the gait data for further analysis[24]. The study takes specific biomechanical and musculoskeletal factors into account when analysing gait data and predicting foot typology using advanced machine learning techniques[25].

2.1 Data collection

The data collection employed a structured approach in gathering data, using various methods such as surveys, informed consent from the participants, body composition analysis, and gait analysis using instrumented treadmill, Zebris, Germany. As illustrated in Fig. 2, this wide range of techniques made sure that multiple parameters were collected resulting in a robust dataset for further study.

Participant selection based on appropriate inclusion and exclusion criteria were followed to ensure validity and applicability of the findings. Participants in this study were divided into two age groups Group 1 (Age 20-29) and Group 2 (Age 40-49) and were exclusively female. All participants were free from significant lower limb abnormalities, long term medication use, recent surgeries, or any history of lower limb injuries. Table 1 provides a list of inclusion and exclusion criteria followed during the study.

Informed consent played a vital role in the study. Participants were briefed on the study goals, methods, risks, and benefits. To maintain the validity and integrity of the study outcomes, participants are carefully examined using assessment questionnaires that included personal information, medical history, and current health condition. This guarantees that only those who have met all selection criteria are included in the observation.

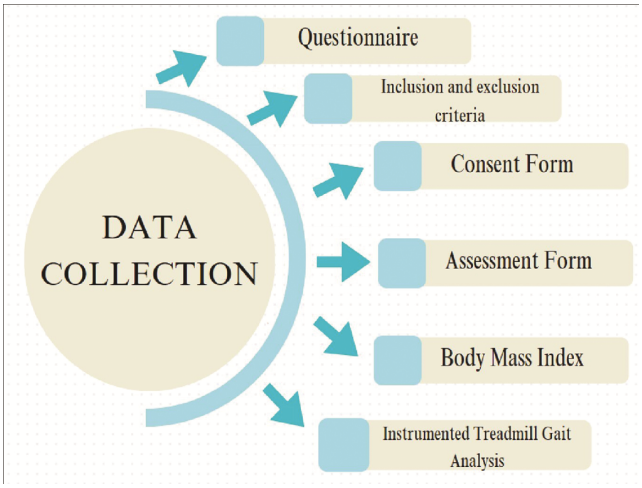

Fig. 2 — Data collection.

Subjects were evaluated using a body composition analyser (IOI 353) and instrumental treadmill gait analysis (ITGA). Bioelectrical Impedance Analysis (BIA) is used by the IOI 353 Body Composition Scale to measure important factors such height, weight, lean body mass, total body water, body mass index (BMI), and basal metabolic rate (BMR). The participants stood barefoot while holding four pole electrodes to determine the body impendence providing detailed information on body composition data.

The ITGA accurately evaluates spatial-temporal variables such as stride length, cadence etc and foot pressure distribution by accurately measuring gait mechanics during static and dynamic conditions. In the static analysis several parameters are investigated such as the foot pressure distribution, postural stability, symmetry of weight distribution, and Centre of Pressure (CoP). For dynamic analysis a camera recorded the participants position on the treadmill during motion. During the 30 second measurement, data for 30 steps per leg (left and right) is collected. Both the analysis is conducted in barefoot condition. The process of collecting data with the instrumented treadmill gait analysis instrument is shown in Fig. 3.

In ITGA, the embedded pressure sensors within the running belt of the treadmill captured real-time data on foot pressure and movement. These signals are then transmitted to a central processing unit where advanced algorithms analyse the data converting raw sensor data into meaningful metrics. For static analysis, sensors measured the distribution of pressure across different regions of the foot while in the dynamic analysis, the sensors record the continuous motion of each footstep synchronized with video footage for gait assessment. This combination of

Table 1 — Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Participants within a specific age range (20-29 years) and (40-49 years) relevant to the target population for gait analysis.	Participants with extreme lower limb injuries that might significantly affect gait.
Participants with diverse backgrounds including various occupations and lifestyles.	Participants with recent major lower limb surgeries or conditions (within 1 year).
Participants who regularly wear various types of footwear (including any specialized footwear or orthotics) considering the impact of different footwear on gait.	Participants who are under medication for a long period of time.
	Participants with acute/chronic lower limb abnormalities or health concerns.

Fig. 3 — Data collection using ITGA.

numerical data and visual analysis provide a thorough evaluation of the participants' gait biomechanics.

2.2 Dataset description

The study involved 40 subjects to participate in the study, with 25 subjects in Group 1 and 15 subjects in Group 2. Each subject underwent three trials in the instrumented treadmill gait analysis (ITGA), where data from both left and right feet were recorded.

For each subject in Group 1, the data from each trial were divided into left and right foot readings, comprising 49 features extracted from images of maximum 30 steps. A dataset consisting of 294 parameters (3 trials * 2 feet * 49 features) per subject is created. Out of the 25 subjects n Group 1, 13 subjects had Normal Arch, 10 subjects had High Arch and 2 subjects had Flatfoot. In total the dataset comprises of 150 samples and each sample represents data from one foot during a trial. For Group 1, the dataset includes 78 samples (13 subjects* 3 trials* 2 feet) for normal arch, 60 samples (10 subjects* 3 trials* 2 feet) for high arch, and 12 samples (2 subjects* 3 trials* 2 feet) for flatfoot. Overall, there are 7,350 (25 subjects * 3 trials * 2 feet * 49 features) parameters for all 25 subjects in Group 1.

Similarly, each subject in Group 2 underwent three trials of ITGA, with readings from both left and right feet. Among the 15 subjects in Group 2, 11 subjects had Normal Arch, resulting in 66 samples (11 subjects* 3 trials* 2 feet) , 2 subjects had High Arch (2 subjects* 3 trials* 2 feet), resulting in 12 samples , and 1 subject had Flatfoot , resulting in 6 samples (1 subjects* 3 trials* 2 feet). In total, there are 90 samples (15 subjects* 3 trials* 2 feet) in Group 2, with each row containing 49 feature columns. Therefore, the entire dataset for Group 2 consists of 4,410 parameters (15 subjects * 3 trials * 2 feet * 49 features).

2.3 Data preprocessing

After completing data collection, the dataset underwent extensive preprocessing to clean and standardize for effective training of a machine learning model. This included handling missing values, normalizing features to ensure consistency, and encoding categorical variables. These steps were crucial to prepare the data for effective training and to improve the model's performance[26]. Following that, the dataset was labelled into different classes. However, due to class imbalance, the SMOTE technique was employed for augmentation. An unbalanced dataset can lead to model bias toward the majority class, resulting in poor performance on minority classes. Hence the SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the class distribution by creating synthetic samples for the minority class[27-29]. By applying SMOTE, the dataset achieved a more balanced class distribution, which is essential for training a robust machine learning model.

2.4 Dataset description after augmentation

For Group 1 and Group 2, the dataset was augmented such that all three-foot typology classes (Normal Arch, High Arch, and Flatfoot) were balanced, resulting in equal data points for each class. After applying SMOTE, Group 1 achieved a balanced dataset with each foot typology class augmented to 78 samples each, totalling 234 samples across all classes (78 samples * 3 classes). This resulted in 11,466 parameters (78 samples * 3 classes * 49 feature parameters per subject). Similarly, in Group 2, each foot typology class was augmented to 66 rows, resulting in a total of 198 samples (66 samples * 3 classes) and 9,703 parameters (66 samples * 3 classes * 49 feature parameters per subject). Therefore, balanced datasets were created for both Group 1 and Group 2 for further training using ML algorithms.

2.5 Feature description

For each sample in the dataset, 49 features were extracted. The features were categorized into spatiotemporal parameters, timing, butterfly parameters, force parameters and pressure parameters. These features as illustrated by Fig. 4 provide a comprehensive foundation for uncovering patterns, understanding sample dynamics, and developing robust predictive models.

A Z-test as illustrated by Equation 1 was performed to identify the most significant features among the

49 features. The significant features had a P-value less than 0.001, indicating a high level of statistical significance.

$$z = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}} \quad \ldots(1)$$

where,

i and j are the indices for the classes being compared.

μ_i and μ_j :: The mean of the i-th or j-th class. It represents the average value of a specific feature for the samples.

σ_i^2 and σ_j^2 :The variance of the i-th or j-th class. It measures the spread or dispersion of the feature values within the group.

n_i and n_j :The sample size of the i-th or j-th class or sample. It represents the number of observations in the group.

The 15 most significant features, based on the Z scores from Table 2 along with the corresponding box plots are illustrated by Fig. 5 and Fig. 6.

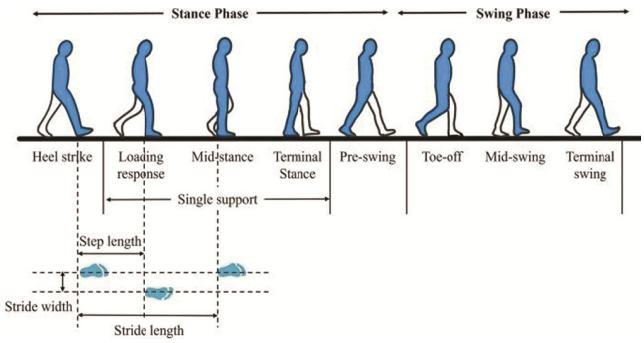

Fig. 4 — Gait cycle[37].

Table 2 — Significant parameters.

Feature	Z Score
Analysis time, sec	2623.55
Contact time Forefoot, % of stance time	680.02
Stance phase, %	423.39
Contact time Midfoot, % of stance time	333.46
Time maximum force Forefoot, % of stance time	276.65
Time maximum force2, %	270.82
Mid stance, %	213.55
Ant/post position, mm	205.80
Swing phase, %	201.06
Cadence, steps/min	195.55
Stride time, sec	195.52
Step time, sec	188.88
Length of gait line, mm	161.45
Total force, %	145.07
Double stance phase, %	118.73

2.6 ML algorithm

The research focuses on classifying foot typology through gait analysis, using machine learning techniques based on their gait parameters. Foot typology is categorized into three main classes: Normal Arch, High Arch, and Flatfoot. Several machine learning models were employed for this classification task, including Decision Tree, Support Vector Machine (SVM), Random Forest, Bagging, Boosting, and K-Nearest Neighbours (KNN)[30]. The models are selected based on the ability to handle multi-class classification from complex datasets.

Real time predictions based on adjacent data points is provided by K-Nearest Neighbour making it suitable for classification[31]. A straightforward interpretability is presented by decision tree machine learning model by classifying data based on features, although they can sometimes overfit[32]. Multiples decision trees are combined to form the Random Forest that is accurate and robust in classification[33]. SVM excels in high-dimensional spaces, using hyperplanes and kernel methods for nonlinear data[34]. Bagging reduces variance and overfitting by training multiple models on different data subsets[35]. Boosting trains the data in loops to correct previous errors, significantly improving accuracy[36]. Hence for classification, the dataset is divided into 70% for training and 30% for testing, following a supervised learning approach with labelled multi-class data. The strategy was chosen due to discrete nature of the dataset, where each individual falls into one of the predefined foot typology categories.

3 Results and Discussion

The results demonstrate the efficiency of machine learning models in classifying foot typology across two distinct age groups: Group 1 (20-29 years) and Group 2 (40-49 years). The Synthetic Minority Over-Sampling Technique (SMOTE) to balance the dataset proved crucial in the data preprocessing step. SMOTE enhanced the dataset, leading to heightened predictive accuracy, precision, recall, and F1 scores across a range of machine learning models. The performance metrics across different ML algorithms for Group 1 and Group 2 respectively are shown in Table 3.

The Bagging algorithm is the most successful machine learning model for categorizing foot typology, according to the results for Group 1 (Age 20–29). The dataset comprises 165 samples for training and 69 samples for testing. The detailed confusion matrix for bagging algorithm

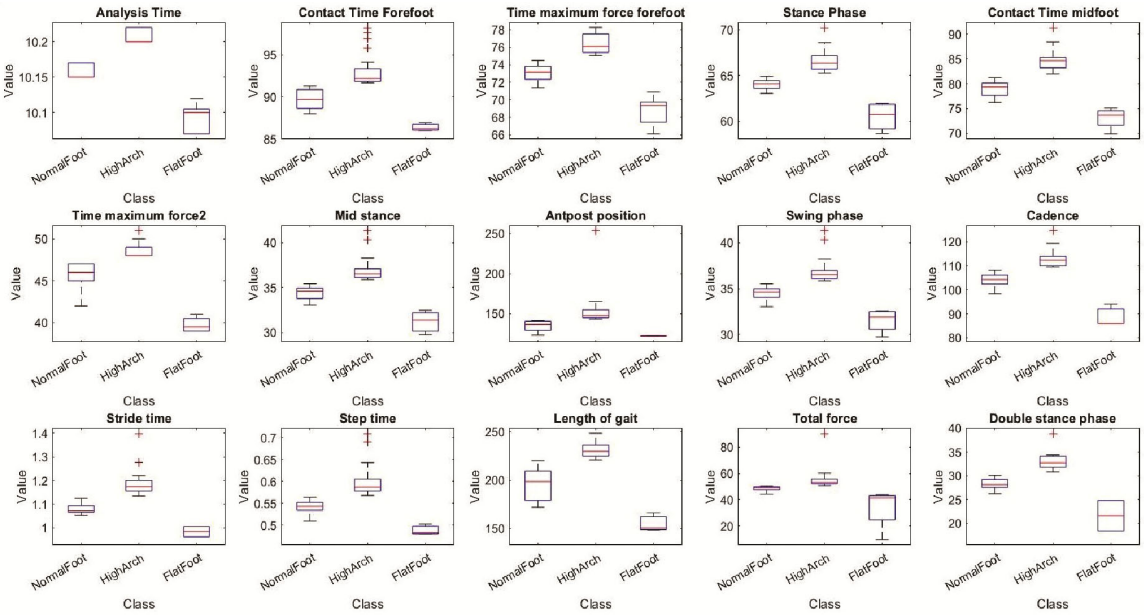

Fig. 5 — Box plot of significant features in Group 1.

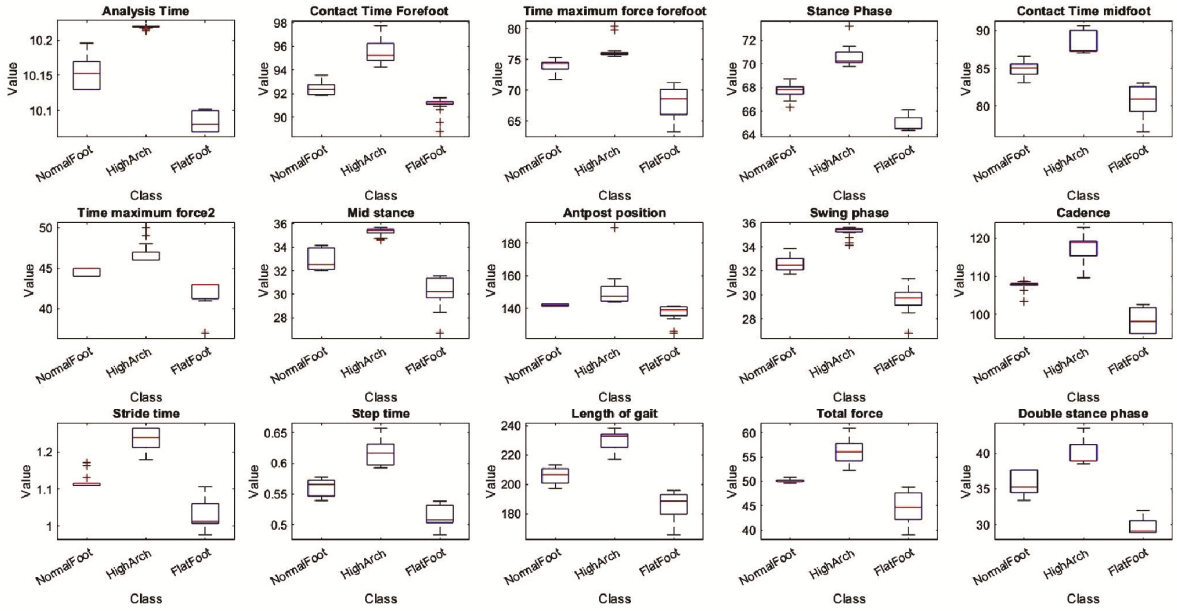

Fig. 6 — Box plot of significant features in Group 2.

Table 3 — Group 2 machine learning performance metrics.

ML Algorithm	KNN	Random Forest	Decision tree	SVM	Bagging	Boosting
Accuracy	91.3	93.2	95	78.2	95.6	89.8
Precision	91.8	93.1	95.7	82.8	95.7	90.2
Recall	91.3	93.8	95.6	78.2	95.6	89.2
F1 Score	91.2	93.5	95.6	77.8	95.6	89.9
			After Augmentation			
Accuracy	92	91	83	61.4	94.3	71.9
Precision	93	93	85	82.1	96.5	76.2
Recall	92	92	83	61.4	96.4	71.9
F1 score	92	92	83	53.8	96.4	72.7

Table 4 — Confusion matrix of bagging - Group 1.

		Predicted		
N=69		Normal	High arch	Flatfoot
Actual	Normal	22	1	0
	High arch	2	21	0
	Flatfoot	0	0	23

Table 5 — Confusion matrix of bagging - Group 2.

		Predicted		
N=57		Normal	High arch	Flatfoot
Actual	Normal	18	1	0
	High Arch	0	19	0
	Flatfoot	1	0	18

represented by Table 4 provides a comprehensive breakdown of classification results, illustrating the number of true positives, true negatives, false positives, and false negatives across the different categories.

Similarly for Group 2, the bagging classifier exhibited good performance in terms of accuracy in categorizing foot typology. The dataset comprised of 141 samples for training and 57 samples for testing. The confusion matrix as shown by Table 5 presents a summary of classification outcomes.

The analysis of machine learning models before and after data augmentation shows significant performance improvements in most models for both age groups. For Group 1, models like KNN, Decision Tree, and Bagging achieved high accuracy improvements, with Decision Tree and Bagging reaching up to 95.6% accuracy post-augmentation. Similarly in Group 2, Bagging exhibited the highest improvement, achieving 94.3% accuracy after augmentation.

4 Conclusion

A three-class classification study was performed to categorize foot typology using machine learning (ML) techniques. Gait parameters from both Group 1 and Group 2 were analysed , employing ML models such as K-Nearest Neighbours (KNN), Random Forest, Bagging, and others to distinguish between different foot typology. The dataset imbalance was addressed and enhanced using Synthetic Minority Over-sampling Technique (SMOTE) to balance class distributions. The models were evaluated with a 70% training and 30% testing data split to ensure reliable model generalization and performance. In Group 1 and Group 2, ML algorithms like KNN, Random Forest, Decision Tree, Bagging, and Boosting demonstrated strong performance, with Bagging algorithm exhibiting high range of accuracy of 95.6 % and 94.3% respectively. In conclusion, the study highlights the integration of ML algorithms with conventional methods to enable detailed analysis of dynamic gait parameters, offering profound insights into foot health and biomechanics. This integrated approach not only enhances diagnostic accuracy but also enables the development of personalized early intervention strategies tailored to specific foot conditions.

Acknowledgments

Authors thank CSIR- CLRI for the facilities used for gait data collection and all the volunteers who had participated in the study. CSIR- CLRI Communication No. 2050.

References

1 Charles J P, Grant B, D'Août K & Bates K T, *J Hum Evol*, 156 (2021) 103014.
2 Holowka N B & Lieberman D E, *J Exp Biol*, 221 (2018) jeb174425.
3 Safavi P S, Janney C, Jupiter D, Kunzler D, Bui R & Panchbhavi V K, *Foot Ankle Spec*, 11 (2018) 193864001880374.
4 Pisal S N, Chotai K & Patil S, *Indian J Forensic Med Toxicol*, 14 (2020) 653.
5 Das P, Jeyakumar S, Thomas A & Lendghar P, *Indian J Physiother Occup Ther*, 18 (2024) 942.
6 Marencakova J, Maly T, Sugimoto D, Gryc T & Zahalka F, *PLoS One*, 13 (2018).
7 Kim D, Lewis C L & Gill S V, *PLoS One*, 16 (2021).
8 Woźniacka R, Bac A, Matusik S, Szczygieł E & Ciszek E, *Eur J Pediatr*, 172 (2013) 683.
9 Lamari N & Beighton P, in: Hypermobility in Medical Practice, Springer (Cham), ISBN: 978-3-031-34914-0, 2023, p. 1.
10 Jung D, Mun K R, Yoo S, Jung H & Kim J, *Proc IEEE Eng Med Biol Soc*, 43 (2021) 4559–4565.
11 Mun K R, Chun S, Hong J & Kim J, *Hum Factors*, 61 (2019) 1077.
12 Cen X, *Stiffness-related coupling analysis of the biomechanical functions of the human foot-ankle complex*, Thesis, 2023.
13 Nikolopoulos D & Safos G K (eds), Foot and Ankle Disorders – Pathology and Surgery, BoD–Books on Demand (Norderstedt), ISBN: 978-3754372654, 2023.
14 All Care Foot & Ankle Center How do flat feet affect your overall health?,(2023).
15 Penn Medicine. Flat foot and high arches treatment. (2023)
16 Fan Y, Fan Y, Li Z, Lv C & Luo D, *PLoS One*, 6 (2011).
17 Mandurah Physiotherapy, Flat feet vs high foot arch, (2024).
18 Ayub A, Yale S H & Bibbo C, *Clin Med Res*, 3 (2005) 116.
19 InformedHealth.org (IQWiG), Cologne, 2018.
20 Blauth W, *Z Orthop Ihre Grenzgeb*, 127 (1989) 3.
21 Rampal V & Giuliano F, *Orthop Traumatol Surg Res*, 106 (2020) S115–S123.

22 Puszczałowska-Lizis E, Dąbrowiecki D, Jandziś S & Zak M, *Med Sci Monit*, 25 (2019) 7746–7754.
23 Rao S, Riskowski J L & Hannan M T, *Best Pract Res Clin Rheumatol*, 26 (2012) 345.
24 Joshi A & Patel B, *Orient J Comput Sci Technol*, 13 (2021) 78.
25 Rajpurkar P, Chen E, Banerjee O & Topol E J, *Nat Med*, 28 (2022) 31.
26 Tomar D & Agarwal S, *Int J Database Theory Appl*, 7 (2014) 99.
27 Panigrahy S, *J Electr Syst*, 20 (2024) 804.
28 Elreedy D & Atiya A F, *Inf Sci*, 505 (2019) 32.
29 Dutta P, Paul S & Majumder M, Research Square (2021).
30 Akkur E & Türk F, *J Med Palliat Care*, 4 (2023) 270.
31 Zhang S, *IEEE Trans Knowl Data Eng*, 34 (2022) 4663.
32 García S, Fernández A & Herrera F, *Appl Soft Comput*, 9 (2009) 1304.
33 Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R & Khovanova N, *Biomed Signal Process Control*, 52 (2019) 456.
34 Kecman V, in: Support Vector Machines: Theory and Applications, Springer (Berlin, Heidelberg), ISBN: 978-3540243848, 2005, p. 1.
35 Ghojogh B & Crowley M, The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial (2019)
36 Bentéjac C, Csörgő A & Martínez-Muñoz G, *Artif Intell Rev*, 54 (2021) 1937.
37 Available at: https://images.app.goo.gl/okV53MuCyoHE6QRg8