

ORIGINAL ARTICLE

Open Access



# Identification of sex from footprint dimensions using machine learning: a study on population of Punjab in Pakistan

Muhammad Awais<sup>1</sup>, Faizana Naeem<sup>2</sup>, Nouman Rasool<sup>3,5</sup> and Sajid Mahmood<sup>4\*</sup>

## Abstract

**Background:** Likewise the fingerprints and palm prints, footprints are also helpful in solving a crime puzzle; however, very few studies have been reported targeting the identification of sex-based upon footprint features. Therefore, the present study aims at the identification of sex using footprint features from the population of Punjab, Pakistan. The foot measurements, i.e., toe length ratio, individual toe lengths, foot breadth, and foot index, are used as features for the identification of sex. Footprint samples were collected from 280 volunteers (142 males and 138 females) from all over Punjab (age range 18–50 years). A sex identification method is proposed in this study employing various machine learning algorithms, i.e., Naïve Bayes, J48, Random Forest, Random Tree, and REP Tree, and compared them.

**Results:** The designed model was cross-validated using 10-fold cross-validation. The results demonstrated the varying accuracy of the machine learning algorithms, using different combinations of footprint features. However, the Naïve Bayes algorithm demonstrated an accuracy of 87.8%, for sex identification, using the combination of toe length and foot indexes.

**Conclusions:** It is concluded that by using a combination of toe length and foot indexes and employing the Naïve Bayes algorithm, sex can be identified more accurately as compared to the other methods.

**Keywords:** Footprints, Sex classification, Pakistani population, Naïve Bayes, J48

## Background

In forensic investigations and medico-legal practices, the determination of a person's identity is the first and one of the most important tasks. Forensic analyses are usually performed by examining different parts of the body, like face, teeth, skull, and fingerprints. Routinely, fingerprints and palm prints are used by forensic officials for identification (Uthman et al., 2012). The reason is that a substantial amount of research has been conducted targeting the fingerprints and palm prints for human identification. Like fingerprints and palm prints, footprints are also widely recovered as pieces of evidence from the crime scenes. It is evident from the reported cases that culprits mostly remove their footwear to reduce the noise of walking, hence leaving their foot impressions at the crime scene (Khan and Moorthy, 2013). Thus, the footprint analysis

can provide a great help in the forensic investigations to identify the sex.

The genetics, lifestyle, and climatic factors define the morphology of the human foot. Also the age, body mass index, sex, and population group effects on shape of the foot. If feet are different from one to other individual, then they are based on these factors. Sex-related differences in foot morphology are important especially in the designing of footwear and forensic anthropology. Furthermore, factors like surface variables and planter pressure distribution define the shape and depth of a foot. Various studies have been reported regarding the determination of sex, based upon information related to the footprints with a reasonable degree of accuracy on different populations across the world. The morphological characteristics being used in these studies have shown variation from population to the population for less erroneous estimation of sex in such populations (Khan and Moorthy, 2013; Krauss et al., 2011; Abd-Elazeem and Yousef, 2013).

\* Correspondence: [sajid.mahmood@umt.edu.pk](mailto:sajid.mahmood@umt.edu.pk)

<sup>4</sup>Department of Informatics and Systems, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan  
Full list of author information is available at the end of the article

In 2017, a study was carried out in Uttar Pradesh, India, to evaluate the importance and trust-ability of footprint dimensions in sex, stature, and age estimation. A total of 400 samples were collected from the ages of 10–65. The print was taken bilaterally, for study. A total of 7 measurements were used, out of which, 5 were toe lengths, recorded as T1–T5, and 2 were breadth dimensions for the right and left footprint. The left footprint measurements were greater than that of the right, for both the males and females. It was concluded that there exists a linear correlation between the footprint length and stature of a person for males and females, both. But for sex, there was a partial correlation found from the footprints and the same for the age (Singh and Yadav, 2017).

Another study on Haryanvi Jaat population of Haryana State of India was conducted in 2016 to find out sexual dimorphism using foot and footprint dimensions. A total of 400 samples (200 males and 200 females) were collected bilaterally with ages between 21 and 25 years. Two measurements were taken, i.e., one for length, from toe 1 to the exterior point of the heel of both feet, and the other was breadth, for the footprint. The study concluded that male has a greater footprint length as compared to female and inverse for footprint index (Walia et al., 2016).

Atamturk, 2010 presented a study to estimate sex from footprint measurements, e.g., heel breadth, ball breadth, and heel ball index and reported that all measurements (length, breadth, and heel breadth) of the footprints were larger in males, but the sex differences were not statistically significant. The study concluded that footprint parameters can be used in the estimation of sex but the heel ball index may not be that much useful in sex determination from footprints (Atamturk, 2010), and extension of this work was reported by (Krishan et al., 2011).

In 2006, a study was conducted on footprints of 320 volunteers for sex estimation in Iban Ethnic, East Malaysia. Linear regression equations were employed on the foot length for discriminating males and females from the given population (Oberoi et al., 2006).

A similar study was conducted on the Ghanaian population in 2015 for sex identification with the sample size of 126 students (60 females and 66 males) aged between 18 and 30 years. The lengths of all foot toes, breadth at the ball (BAB), heel ball index (HB index), and the breadth at heel (BAH) were used. The results demonstrated that all footprint dimensions have shown significant results except HB index. The accuracy to discriminate both sex was reported at 80% when toe 5 and BAH from the left footprints were used for analysis (Abledu et al., 2015).

Different parameters such as foot length and size, heel ball index, foot breadth measurements, i.e. breadth at ball and breadth at heel, have been used across the world by many researchers for various populations (Krishan et

al., 2012, 2015). However, sex estimation through feet anthropometry is still an open problem as the scope is very wide due to diversity in the world's population.

No studies have been reported related to the determination of sex-based upon footprint analysis on Pakistani population. The aim of the present study is to find out the utility of footprint for sex classification. For this purpose, footprint sample of the population of Punjab of Pakistan has been collected from different geographical areas of the province. Footprints of 142 male and 138 female volunteers were collected. All toe lengths, ratio of toe lengths, ball breadth index, and heel breadth index of both the feet of each individual from the study were measured from these footprints and analyzed for the sex determination.

Machine learning is a branch of computer science which gives the ability to the system to learn and predict future results with unseen data, it is also referred as the computational statistics to build predictive models on given data and predict for unseen values, without explicit programming (Rughani and Bhatt, 2017). Machine learning is helping the forensics teams across the world in many ways; from individual identification, forensic cyber security, computer forensics, and forensic criminology are used to prevent and solve the crime cases (Ariu et al., 2011; Nasrabadi, 2007).

Machine learning provides different types of computer algorithms to solve various real-world problems, such as Naïve Bayes, Random Forest, Random Tree, REP Tree, and J48 algorithm, particularly for classification problems. Sex identification is also a classification problem; therefore, these machine learning classification algorithms can help in prediction of sex using considered parameters, which can aid forensic analysts in sex identification cases (Brennan and Oliver, 2013; Kim et al., 2014; Nath, 2006; Wang et al., 2013).

In this paper, Naïve Bayes, Random Forest, Random Tree, REP Tree, and J48 machine learning algorithms were used for sex classification. Naïve Bayes is simple but a very powerful classification algorithm based on the Bayesian inference for providing some reasonable accuracy value (Cichosz, 2014). Random Forest, Random Tree, and REP Tree build iterations of trees and reduce the errors by using mean squared, bagging and choosing best respectively for the best classification (Aljawarneh et al., 2017). J48 is a widely used machine learning algorithm for classification with very much enhanced pruning technique and fault tolerance that reduces misclassification (Heude et al., 2005).

## Materials and methods

### Sample collection

The present study was conducted on unrelated adult volunteers randomly selected from Punjab, Pakistan. The volunteers included 142 males and 138 females of age

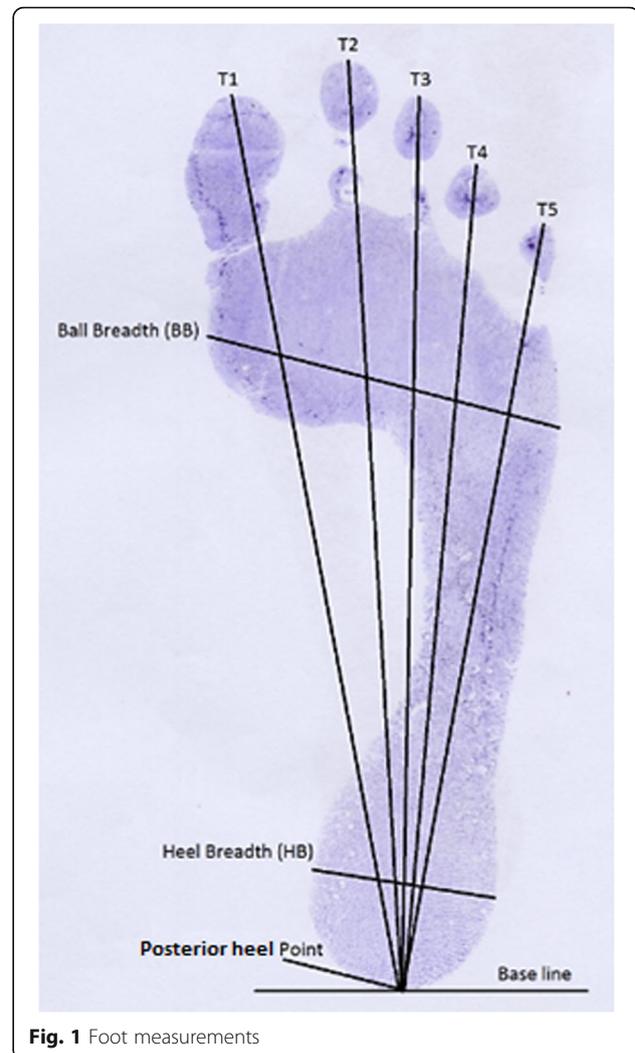
ranging from 18 to 50 years. None of the volunteers had family relations with each other as the anthropometric features of the human body are based on gene, and family dependence can lead to the similarity in the anthropometric features (Heude et al., 2005). The research procedure strictly followed the ethical research standards of University of Management and Technology, Lahore Pakistan.

Before sample collection, all volunteers were properly explained about the purpose of the study. All volunteers were active and healthy with no previous surgical history. To remove dirt, all volunteers were asked to clean their feet with soap and water. The quick-drying duplicating ink was uniformly spread on  $0.30 \times 0.30$ -m plain glass plate of  $0.008 \times 0.008$ -m thickness Rughani and Bhatt, 2017; Singh and Yadav, 2017. The participants were asked to put their feet one by one on the glass plate with normal force and then placed it on a plain A4 size white paper and lift up their foot without disturbing the paper (Robbins, 1986). All papers with vague footprints were excluded from the study, and the participant was requested to repeat the process for a clean print. By using this method, the static footprints of all participants were collected for both feet, i.e., left and right. In this paper, different measurements were used for the identification of sex. The identification can also be carried out if the crime scene investigator does not have all of the measurements in case of dynamic footprints, as all the static features are not usually present in a dynamic footprints (Reel et al., 2012).

#### Method flow chart

The first step of this study was the collection of footprints data from 280 adult volunteers including 138 females and 142 males from different areas of Punjab. The second step was taking the measurements of selected features, i.e., toe lengths, ball breadth, and heel breadth of the feet. A straight line was drawn from the most endpoint of heel to the lateral point of the first toe as shown in (Fig. 1), where the measurements of the length of toe were considered as the distance between the outer most points of the heel to the tip of the toe. Ball breadth was measured as the dimension between the most medial and the most lateral points of the footprint at the ball, and heel breadth was measured as the broadest distance through the heel. Table 1 contains the measurements that have been taken from each footprint. It clearly defines the standard followed in this study that how the lengths of 5 toes, ball breadth, and heel breadth was measured from the feet. The ball breadth index and heel breadth index were calculated using the following formula:

$$\text{BBI:ball breadth index} = (\text{ball breadth/length of toel}) \times 100$$



**Fig. 1** Foot measurements

$$\begin{aligned} \text{HBI : heel breadth index} \\ = (\text{heel breadth/length of toel}) \times 100 \end{aligned}$$

Furthermore, the ratios between all toes were calculated as (toe 1 to toe 2, toe 1 to toe 3, toe 1 to toe 4, toe 1 to toe 5, toe 2 to toe 3, toe 2 to toe 4, toe 2 to toe 5, toe 3 to toe 4, toe 3 to toe 5, and toe 4 to toe 5).

The third step was the implementation of classification techniques for sex identification. Algorithms that were chosen for this purpose were Naïve Bayes, Random Forest, Random Tree, REP Tree, and J48 algorithm (Cichosz, 2014; Kim et al., 2014; Rughani and Bhatt, 2017). The measurements obtained in the second step were the input parameters to the abovementioned algorithms. In the fourth step, a final decision was made by the classification algorithms used (in step 3) for sex identification. The algorithm will use the features of feet and perform its analysis

**Table 1** Parameters and their measurements

N	Measurements
Toe 1	Length from the most endpoint of heel to the most inner point of toe 1.
Toe 2	Length from the most endpoint of heel to the most inner point of toe 2.
Toe 3	Length from the most endpoint of heel to the most inner point of toe 3.
Toe 4	Length from the most endpoint of heel to the most inner point of toe 4.
Toe 5	Length from the most endpoint of heel to the most inner point of toe 5.
Ball breadth (BB)	The dimension between the most medial and the most lateral points of the footprint at the ball (which resembles the most noticeable areas of the metatarsal).
Heel breadth (HB)	The broadest distance through the heel.

for differentiating males from females in the available sample (Fig. 2).

**Statistical analysis**

The statistical analysis was performed using Weka 3.8 for windows. We employed Naïve Bayes, J48, Random Forest, Random Tree, and REP Tree for sex classification using BBI and HBI (Aljawarneh et al., 2017; Cichosz, 2014; Kalmegh, 2015). Naïve Bayes is a classifier based on Naïve Bayes theorem. It requires a small set of training data to identify the sex. Random Forest classifier fits a number of decision tree classifiers on different samples of the dataset and calculates the average of all in order to improve the accuracy.

Random Tree, randomly built trees; no pruning is involved and performs analysis on the basis of these trees. REP Tree is a fast decision tree learning algorithm. It is based on information gain and entropy which minimizes the error that arises due to variance. It considers all the attributes and constructs the decision tree with the help of variance and information gain and does error pruning. J48 is widely used as it creates a binary tree, and after building the tree, the algorithm is applied in each record of the database and classification is performed.

Naïve Bayes algorithm for continuous data is based on the equation

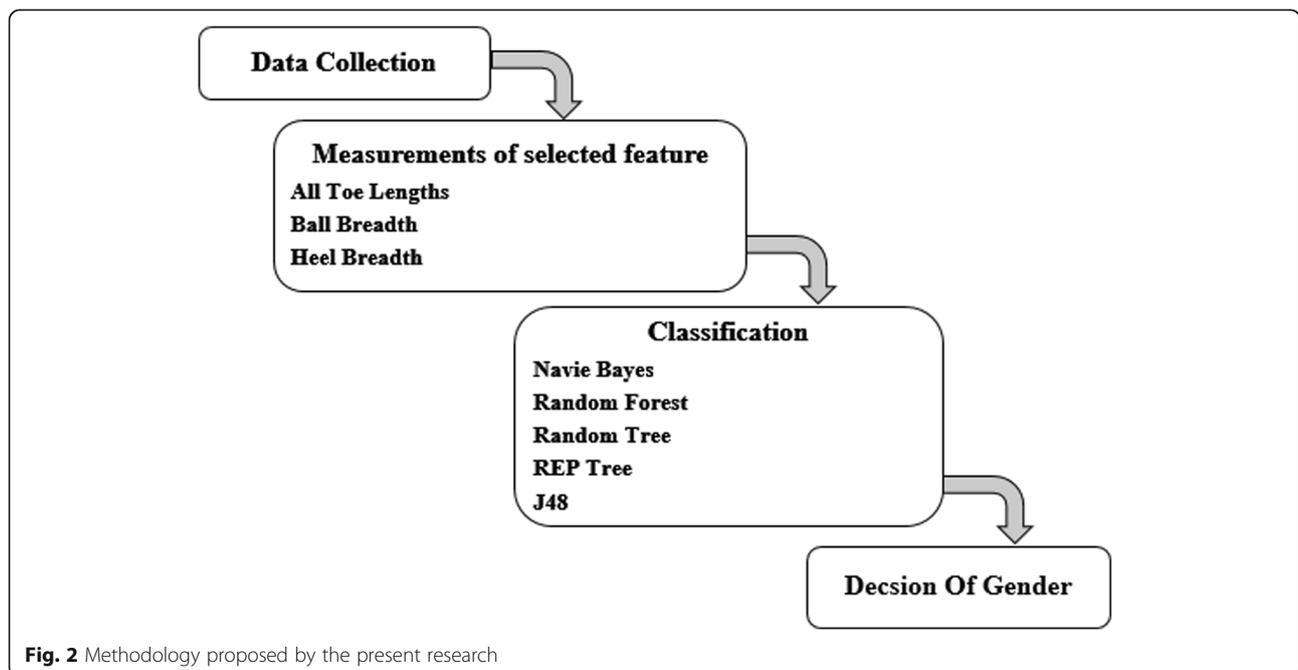
$$p(x = v \vee C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{1}$$

Let  $x$  be the length of a given toe, where data segments on male and female classes and variance are calculated for each male and female class, and let  $\mu_k$  be the mean value of  $C_k$ , i.e., associated class and  $\sigma_k^2$  the variance of associated class  $C_k$  and  $v$  the observation with unknown class, i.e., either male or female. Then, for a given class, the probability distribution  $C_k$ ,  $p(x = v \vee C_k)$  can be computed by putting values into the equation.

Random Tree algorithm is based on

$$f = \frac{1}{B} \sum_{b=1}^B f_b(x') \tag{2}$$

The J48 algorithm uses the entropy and information gain for making a decision for determination of sex.



**Fig. 2** Methodology proposed by the present research

$$H(T) = \sum_{i=1}^j p_i \log_2 p_i \tag{3}$$

$$IG(T, a) = H(T) - H(T/a) \tag{4}$$

The mean and standard deviation of the data was computed by using:

$$\mu = \frac{\sum x}{N} \tag{5}$$

$$\alpha = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \tag{6}$$

where  $N$  is the total number of samples (male and female), “ $x$ ” is the sample, “ $\alpha$ ” is the standard deviation, and “ $\mu$ ” is the mean of all samples.

**Results**

In this experiment, we identify the sex using different foot parameters (foot toe lengths, ratios of all foot toe lengths, ball breadth index, and heel breadth index). After performing the analysis, we attained the following results.

Abbreviation, i.e., LFT1, for left foot toe 1 up to LFT5 for left foot toe 5 and similarly RFT1, RFT2...RFT5 for right foot toe 1, was used in the description of the results.

**Comparison between the mean of right and left foot (both males and females)**

In males, a greater mean value was observed for foot toe 1 and ball breadth index of the left foot, while a higher mean value was observed for the heel breadth index of the right foot in comparison to the left foot.

Similarly, in females, a greater mean value was observed for ball breadth index and foot toe 1 of the left foot, whereas a higher mean value of heel breadth index was observed of the right foot as compared to left (Table 2).

**Comparison between the standard deviation of the right and left foot (both males and females)**

In males, the standard deviation was noticeably high for the foot toe 1 and heel breadth index of the right foot, while a higher standard deviation was noticed for the ball breadth index of the left foot as compared to right.

In females, the higher standard deviation was observed in foot toe 1 and ball breadth index of the right foot, whereas a higher standard deviation value was noticed for the heel breadth index of the left foot in comparison to the right (Tables 2 and 3).

**Determination of accuracy**

The accuracy in the classification problems is measured as how much the algorithm identifies a correct class for the given instance of data and whereas the error is the number of percent of misclassified instances of data (Ariu et al., 2011). In this paper, there were two classes: “male” and “female”, the precision of the algorithm will be measured as how much the machine learning algorithm correctly classified a male as male and a female as female.

**Sex classification based on left foot parameters (toe 1 length and indexes)**

Sex classification based on the left foot ridges was performed using different algorithms (Naïve Bayes, J48, Random Forest, Random Tree, and REP Tree). The input parameters to these algorithms include the LFT1 (left foot toe 1), BBI, and HBI of the left foot.

Using the J48 algorithm, the value of TPs (true positives/accuracy) was higher in males as compared to females. The accuracy rates were 0.873 and 0.841 for males and females, respectively, while the overall accuracy rate was observed at 85.7% (Table 4).

The Random Forest algorithm demonstrated an overall accuracy rate of 85%, while the TPs (accuracy) for females were recorded at 84.1 and 85.9% for males. By applying Random Tree algorithm, the overall accuracy rate was observed at 83.9%. However, the value of TPs was greater for females than that of males. Individually, the accuracy rates were 0.841 and 0.838 for females and males, respectively (Table 4).

The REP Tree algorithm exhibited an accuracy rate of 82.5%, whereas the TP values were 0.826 and 0.824 for females and males, respectively. Through the Naïve Bayes algorithm, the accuracy rate was 0.906 for females and 0.852 for males, whereas the overall accuracy rate of sex classification was observed at 87.8% (Table 4).

Hence, if we use FT1, BBI, and HBI of the left foot as parameters, the Naïve Bayes algorithm is more suitable

**Table 2** Descriptive statistics: left foot dimensions and foot index among Punjabi population

	Foot toe 1		Ball breadth index		Heel breadth index	
	Male	Female	Male	Female	Male	Female
N	142	138	142	138	142	138
Mean	24.0	21.7	38.2	38.2	20.1	20.2
SD	1.2	0.8	1.9	2.0	1.6	1.9
Range	21.4–27.9	19.5–23.9	34.03–43.90	33.0–43.1	13.53–24.54	15.57–25.15

**Table 3** Descriptive statistics: right foot dimensions and foot index among Punjabi population

	Foot toe 1		Ball breadth index		Heel breadth index	
	Male	Female	Male	Female	Male	Female
N	142	138	142	138	142	138
Mean	23.9	21.7	38.2	37.9	20.2	20.3
SD	1.2	0.8	1.9	2.1	1.7	1.7
Range	20.58–27.65	19.6–24.26	33.75–44.73	32.1–44.1	16.55–25.21	15.19–24.37

for sex determination with 87.8% accuracy. However, the REP Tree algorithm demonstrated the lowest accuracy rate (82.5%).

#### Sex classification based on left foot parameters (toe lengths and ratios)

Sex classification was performed using left foot parameters (LFT1+ LFT2+ LFT3+ LFT4+ LFT5+ left foot toe ratios). Through Random Tree, Random Forest, and Naïve Bayes algorithm, the value of TPs was observed higher in females than that of males. In comparison, through J48 and REP Tree algorithm, the value of TPs was observed higher in males than that of females.

Through the Naïve Bayes algorithm, the accuracy rate was 0.91 for females and 0.83 for males while the overall accuracy was observed at 87.5% (Table 5). The accuracy rates, using Random Tree algorithm, were 0.833, 0.824 for females and males respectively while the collective accuracy rate was 82.8%. The accuracy rates, using the Random Forest algorithm, were 0.870 and 0.859 for females and males, respectively. However, the collective accuracy rate was 86.4%. The J48 algorithm showed an accuracy rate of 0.859 for males and 0.855 for females. However, the overall accuracy rate was at 85.7% (Table 5). The accuracy rates obtained by using REP Tree algorithm were 0.873 and 0.841 for males and females, respectively, while the collective accuracy rate was 85.7% (Table 5).

**Table 4** Accuracy rates for sex determination through following algorithms using toe length and foot indexes

Algorithm	Sex	TP	Accuracy (%)
J48	F	0.841	85.7
	M	0.873	
Random Forest	F	0.841	85
	M	0.859	
Random Tree	F	0.841	83.9
	M	0.838	
REP Tree	F	0.826	82.5
	M	0.824	
Naïve Bayes	F	0.906	87.8
	M	0.852	

Thus, Naïve Bayes algorithm demonstrated the highest rate of accuracy at 87.5% for sex classification by using the selected parameters of the left foot, whereas Random Tree algorithm showed the lowest accuracy rate of 82.8% among all.

#### Sex classification based on right foot parameters (toe 1 length and indexes)

In this experiment, we identify the sex through J48, Random Tree, Random Forest, REP Tree, and the Naïve Bayes algorithms using right foot parameters (FT1, HBI, and BBI). Using these algorithms, it was observed that, in all of these, the value of TPs (accuracy) was greater in females than that of males while through REP Tree, males showed greater TP value. The accuracy rate by using the J48 algorithm was 0.88 and 0.80 for females and males, respectively. However, the overall accuracy rate was observed at 84.2% (Table 6). The Random Tree algorithm revealed an accuracy of 0.812 for females and 0.810 for males. However, their combined accuracy was observed at 81% (Table 6). Naïve Bayes algorithm demonstrated that there is a significant difference between the TP value of both sex. The accuracy rate was 0.92 in females and 0.81 in males. Through the Naïve Bayes algorithm, the overall accuracy rate of sex classification was observed at 86.4% (Table 6). The overall accuracy rate by using Random Forest algorithm was recorded at 84.2%. It showed an accuracy of 0.877 for females and 0.810 for males (Table 6).

**Table 5** Accuracy rates for sex determination through algorithms using toe lengths and toe length ratios

Algorithm	Sex	TP	Accuracy (%)
J48	F	0.855	85.7
	M	0.859	
Random Forest	F	0.870	86.4
	M	0.859	
Random Tree	F	0.833	82.8
	M	0.824	
REP Tree	F	0.841	85.7
	M	0.873	
Naïve Bayes	F	0.913	87.5
	M	0.838	

**Table 6** Accuracy rates for sex determination through following algorithms by using toe 1 lengths and indexes

Algorithm	Sex	TP	Accuracy (%)
J48	F	0.884	84.2
	M	0.803	
Random Forest	F	0.877	84.2
	M	0.810	
Random Tree	F	0.812	81
	M	0.810	
REP Tree	F	0.848	85.3
	M	0.859	
Naïve Bayes	F	0.920	86.4
	M	0.810	

Through REP Tree algorithm, a greater TP value was observed for males than that of females. The accuracy rates were 0.859 and 0.848 for males and females, respectively. Collectively, the accuracy rate was observed at 85.3% (Table 6). Now, we can say that if we use the parameters of the right foot (FT1, BBI, and HBI), then the Naïve Bayes algorithm gives the best accuracy rate (86.4%) among all. That is why it is the most suitable choice for these parameters, whereas Random Tree Algorithm gives the lowest accuracy rate of 81%.

#### Sex classification based on right foot parameters (toe lengths and ratios)

Sex classification was performed using the right foot parameters (RFT1+ RFT2+ RFT3+ RFT4+ RFT5+ right foot toe ratios). Through J48, Random Forest, REP Tree, and Naïve Bayes algorithms, TP values observed were greater in females as compared to males.

By the J48 algorithm, the value of TPs was 0.819 and 0.817 for females and males, respectively, while the overall accuracy was 81.7%. The accuracy rates by using Random Forest algorithm were 0.870 and 0.845 for females and males, respectively, while the overall accuracy rate was observed at 85.7% (Table 7). By REP Tree algorithm, the value of TPs was greater for females in comparison to males. The accuracy rate was observed at 0.906 and 0.810 for females and males, respectively. However, the overall accuracy rate was observed at 85.7% (Table 7). The Naïve Bayes algorithm showed an accuracy rate of 0.964 for females and 0.68 for males, whereas the overall accuracy rate was observed at 82.1% (Table 7). By applying Random Tree algorithm, the value of TPs was greater in males than that of females. The accuracy rates were 0.859 and 0.797 for males and females, respectively, while the overall accuracy rate was observed at 82.8% (Table 7).

Thus, if we use the selected parameters of the right foot and perform the sex classification analyses, Random Forest and REP Tree algorithms demonstrate the best

**Table 7** Accuracy rates for sex determination through following algorithms by using toe lengths and toe length ratios

Algorithm	Sex	TP	Accuracy (%)
J48	F	0.819	81.7
	M	0.817	
Random Forest	F	0.870	85.7
	M	0.845	
Random Tree	F	0.797	82.8
	M	0.859	
REP Tree	F	0.906	85.7
	M	0.810	
Naïve Bayes	F	0.964	82.1
	M	0.683	

accuracy rates (85.7%) among all, whereas J48 revealed the lowest accuracy rate (81.7%) as in comparison to all the other algorithms.

#### Overall result

After observing all the results, it can be concluded that the accuracy rates for identifying sex vary with the changing of parameters as described in Tables 4, 5, 6, and 7. Naïve Bayes algorithm showed the highest accuracy rates for all type of parameters except the ratios and toe lengths of the right foot (Table 7). It shows accuracy rates of 87.8, 87.5, 86.4, and 82.1% for the parameters used in Tables 4, 5, 6, and 7, respectively.

Naïve Bayes algorithm performed best among all algorithms with an overall accuracy rate of 86%. It can also be observed from these results that Random Tree algorithm has not performed as better as the other algorithms applied in this study. It demonstrated the lowest accuracy rates of 83.9, 82.8, 81, and 82.8% in Tables 4, 5, 6, and 7, respectively, and overall 82.6% accuracy.

#### Discussion

Across the world, many researchers have taken the initiative for utilization of footprints in sex determination. Different features of the foot have been used in various studies for accurate identification of sex. No such study has been conducted on any of the Pakistani population to date. The present study was conducted on the population of Punjab using different foot measurements, i.e., foot toe lengths, indexes, and toe ratios, for sex identification. All these parameters, i.e., foot toe lengths, indexes, and toe ratios for the left foot have greater values as compared to those of the right foot. The maximum accuracy rate of determination of sex was observed at 87.8% for the left foot and 86.4% for the right foot.

Previously, many reports have been published on foot-related studies and on the individuality of the foot and footprints, sexual dimorphism, height estimation, and

association of foot and hand dimensions for the identification of a person in mass disasters (Kanchan et al., 2008; Krishan et al., 2011) (Table 8).

A study on Ghanaian population has shown that by using toe 1 (length of the toe 1), toe 2, toe 3, toe 4, toe 5, BAH (breadth at heel), and BAB (breadth at ball), the reasonable accuracy 69.8–80.3% can be obtained for determination of sex (Abledu et al., 2015). However, in the present study, several parameters have been used, i.e., toe 1, toe 2, toe 3, toe 4, toe 5, ratios of the toes, ball breadth index, and heel breadth index. Different algorithms have been used for classification of sex, but Naïve Bayes algorithm has the highest performance (87.8%) in all. If we compare the results of the present study with the study performed on the Ghanaian population, it can be seen that the results are better by the methodology proposed in this study (Table 8).

The results of this study can also be compared with the results described by Krishan, et al. in 2012, an Indian population-based study (Krishan et al., 2012). Heel ball index and ball breadth index of both feet have been used, and it was observed that the measurements at the heel and ball were significantly greater in males than females. They have concluded that heel ball index is not a significant parameter in sex determination (Krishan et al., 2011). But in the present study, mean value of ball breadth index is slightly greater in males, and the mean value of HBI is slightly higher in females (0.02 and 0.04 for left and right foot, respectively). The results of sex determination with these two parameters were not highly significant (almost 70%). Therefore, to increase the accuracy rate of sex determination, length of toe 1 was used along with the indexes. A significant rate of 87.8% accuracy was observed by the combination of these three parameters.

The table below provides a comparison of different studies that have been conducted previously for the discrimination of sex using footprints. It is evident from the above table that the method proposed in the current study outperforms those reported in the literature. Only the method proposed in (Abd-Elazeem and Yousef, 2013) has the same accuracy, i.e., 88%. However, they have used only two parameters from footprint

measurements (foot length and foot breadth). Also, their experimentation was conducted on a sample with a small age bracket, i.e., 20–35 years of age (Abd-Elazeem and Yousef, 2013) whereas our study is based on the four footprint measurement parameters (lengths of all toes, ball breadth, heel breadth, and all foot toe ratios). In addition, the age window is much larger, i.e., from 18 to 50-year-old people which makes the results more robust and less biased. All other reported studies have lesser accuracy as well as a lesser number of parameters and smaller age windows. A bigger age window is helpful in the forensic analysis as the crime perpetrators are generally in the age bracket of 18–50 years.

After comparing the proposed method with others available for sex determination, it can be said that if we use foot length, heel breadth index, and ball breadth index and perform sex classification using Naïve Bayes algorithm, then the sex identification could be achieved with more precision and accuracy. Therefore, it will be feasible to use these parameters for sex determination for the population of Pakistan.

## Conclusion

This study demonstrates a reasonable degree of association between the footprint and sex determination from the footprint samples of 280 volunteers of both sexes aged between 18 and 50 years. Different classification algorithms, i.e., J48, Random Tree, Random Forest, REP Tree, and Naïve Bayes, have been used for the classification of sex. The results obtained from the Naïve Bayes algorithm found to be more accurate in predicting sex compared to the other algorithms. The percentage accuracy of establishing sex using ball breadth index, heel breadth index, and the toe 1 length is 87.8% with the help of Naïve Bayes algorithm, which is quite significant for use, the results are better by using left footprint parameters. It is possible that selected footprint measurements that show significant results for one population may not demonstrate the significant results for the other one. Therefore, more research needs to be conducted on different populations to determine sex from the foot with reliable accuracy rate. This type of study is conducted for

**Table 8** Comparison of the present study and other reported studies

Study	Population	Age-group	Sample size	Parameters	Accuracy (%)
(Sen et al., 2011)	Rajbanshi, India	18–50	300	Foot breadth and foot length	80
(Oberoi et al., 2006)	Mangalore, India	18–26	200	Foot length	80
(MoustafaAbd-Elazeem and Yousef, 2013)	Fayoum, Egypt	20–35	200	Foot breadth and foot length	88
(Hemy et al., 2013)	Western Australian	18–68	200	All toe length, ball breadth and heel breadth	81
(Krishan et al., 2015)	North Indian	13–18	303	All foot toes' ratios	60
Present study	Punjab, Pakistan	18–50	280	All toe length, ball breadth, heel breadth and all foot toes' ratios	88

the first time on individuals of the population of Punjab, Pakistan. The important application of this study is the identification of an individual during forensic investigations based on footprint measurements. Further studies with large data samples from different geographical areas of Pakistan could improve the utility of this kind of research.

#### Acknowledgements

We would like to thank our colleagues because this work could not be done without the support of our colleagues Mr. Mashal Khalid and Mr. Abdulrehman Asghar. We are grateful to them for their efforts and for sharing their pearl of wisdom with us that improve this manuscript. We would also like to thank all those that cannot be mentioned here for their supports especially to all the volunteers that were part of this study.

#### Funding

No funding received for this study.

#### Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

#### Authors' contributions

MA and FN collected and analyzed the data. NR performed the interpretation of results and was a major contributor in writing the manuscript. SM supervised the whole study and helped in discussion writing for the results. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

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

#### Author details

<sup>1</sup>Department of Information System, School of Business and Economics, University of Management and Technology, Lahore, Pakistan. <sup>2</sup>Department of Software Engineering, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan. <sup>3</sup>Department of Life Sciences, School of Science, University of Management and Technology, Lahore, Pakistan. <sup>4</sup>Department of Informatics and Systems, School of Systems and Technology, University of Management and Technology, Lahore, Pakistan. <sup>5</sup>Dr Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan.

Received: 21 September 2018 Accepted: 5 December 2018

Published online: 22 December 2018

#### References

- Abd-Elazeem, G, Yousef, W (2013). Sex Determination using Finger Length, Fingers Ratio, and Foot Dimensions in Fayoum Governorate. *Ain Shams Journal of Forensic Medicine and Clinical Toxicology*. 21(2);12–19. <https://doi.org/10.21608/ajfm.2013.19163>
- Abledu JK, Abledu GK, Offei EB, Antwi EM (2015) Determination of sex from footprint dimensions in a Ghanaian population. *PLoS One* 10:e0139891
- Aljawarneh S, Yassein MB, Aljundi M (2017) An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Clust Comput*:1–17
- Ariu D, Giacinto G, Roli F (2011) Machine learning in computer forensics (and the lessons learned from machine learning in computer security). In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, pp 99–104
- Atamturk D (2010) Estimation of sex from the dimensions of foot, footprints, and shoe. *Anthropol Anz*:21–29
- Brennan T, Oliver WL (2013) The emergence of machine learning techniques in criminology: implications of complexity in our data and in research questions. *Criminol Public Policy* 12:551–562
- Cichosz P (2014) *Data mining algorithms: explained using R*. John Wiley & Sons
- Hemy N, Flavel A, Ishak N-I, Franklin D (2013) Sex estimation using anthropometry of feet and footprints in a Western Australian population. *Forensic Sci Int* 231:402 e401–402. e406
- Heude B, Kettaneh A, Rakotovo R, Bresson J-L, Borys J-M, Ducimetière P, Charles M-A (2005) Anthropometric relationships between parents and children throughout childhood: the Fleurbaix-Laventie Ville Santé Study. *Int J Obes* 29:1222
- Kalmegh S (2015) Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news. *Int J Innov Sci Eng Technol* 2:438–446
- Kanchan T, Menezes RG, Moudgil R, Kaur R, Kotian M, Garg RK (2008) Stature estimation from foot dimensions. *Forensic Sci Int* 179:241 e241–241. e245
- Khan HBMA, Moorthy TN (2013) Stature estimation from the anthropometric measurements of foot outline in adult indigenous Melanau ethnics of East Malaysia by regression analysis. *Sri Lanka Journal of Science & Law*
- Kim G-H, Trimi S, Chung J-H (2014) Big-data applications in the government sector. *Commun ACM* 57:78–85
- Krauss I, Langbein C, Horstmann T, Grau S (2011) Sex-related differences in foot shape of adult Caucasians—a follow-up study focusing on long and short feet. *Ergonomics* 54:294–300
- Krishan K, Kanchan T, Passi N, DiMaggio JA (2012) Heel–ball (HB) index: sexual dimorphism of a new index from foot dimensions. *J Forensic Sci* 57:172–175
- Krishan K, Kanchan T, Passi N, DiMaggio JA (2015) Sexual dimorphism in foot length ratios among North Indian adolescents. *J Forensic Legal Med* 36:96–101
- Krishan K, Kanchan T, Sharma A (2011) Sex determination from hand and foot dimensions in a North Indian population. *J Forensic Sci* 56:453–459
- Nasrabadi NM (2007) Pattern recognition and machine learning. *J Electron Imaging* 16:049901
- Nath SV (2006) Crime pattern detection using data mining. *Web intelligence and intelligent agent technology workshops, 2006 wi-iat 2006 workshops 2006 ieee/wic/acm international conference on*. IEEE, pp 41–44
- Oberoi DV, Kuruvilla A, Saralaya K, Rajeev A, Ashok B, Nagesh K, Rao NG (2006) Estimation of stature and sex from footprint length using regression formulae and standard footprint length formulae respectively. *J Punjab Acad Forensic Med Toxicol* 6:5–8
- Reel S, Rouse S, Obe WW, Doherty P (2012) Estimation of stature from static and dynamic footprints. *Forensic Sci Int* 219:283 e281–283. e285
- Robbins LM (1986) Estimating height and weight from size of footprints. *J Forensic Sci* 31:143–152
- Rughani PH, Bhatt P (2017) Machine learning forensics: a new branch of digital forensics. *Int J Adv Res Comput Sci* 8
- Sen J, Kanchan T, Ghosh S (2011) Sex estimation from foot dimensions in an indigenous Indian population. *J Forensic Sci* 56:S148–S153
- Singh P, Yadav B (2017) Age, sex and stature estimation from footprint dimensions. *Arab J Forensic Sci Forensic Med* 1 (AJFSFM), 1(5).
- Uthman A, Al-Rawi N, Al-Timimi J (2012) Evaluation of foramen magnum in gender determination using helical CT scanning. *Dentomaxillofac Radiol* 41:197–202
- Walia S, Modi BS, Puri N (2016) Sexual dimorphism from foot dimensions and foot prints in Haryana Jat population. *Int J Anat Res* 4:2142–2147
- Wang T, Rudin C, Wagner D, Sevieri R (2013) Learning to detect patterns of crime. *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp 515–530