Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV) &
3rd International Conference on Imaging, Vision & Pattern Recognition (IVPR)

Prediction of Gender and Age from Inertial Sensor-based Gait Dataset

Kanij Mehtanin Khabir‡, Md. Sadman Siraj†, Masud Ahmed, and Mosabber Uddin Ahmed
Department of Electrical and Electronic Engineering, University of Dhaka.
Email: ‡kanij.mehtanin.khabir@ieee.org, †sadman.siraj@ieee.org

***Abstract*—The technology pertaining to gait generation, recognition and analysis is getting more sophisticated each day. The global demand for a gait-based dataset and its subsequent recognition, leading to extraction of valuable information is now higher than ever. However, inertial sensor-based gait dataset is a comparatively new addition to the field of gait analysis. Consequently, most of the research works incorporating machine learning algorithms into gait dataset are image based. In addition to that, most of the gait-based datasets have been analyzed for gait recognition. There remains very little research work on personal authentication from inertial sensor-based gait dataset. Personal authentication is of several types out of which, predicting gender and age is quite challenging. In this paper, we have tried to face these challenges and have manifested the process of predicting gender and age from the inertial sensor-based gait dataset which is a part of the vast Osaka University-ISIR Gait Database. Finally, we have found that the models, Support Vector Machine shows the highest accuracy for the problem of classifying gender and Decision Trees shows the highest variance score (R^2 value) for the problem of predicting age.**

***Index Terms*—gait, inertial sensor, supervised machine learning, gender and age estimation, classification and regression**

I. Introduction

Gait or gait cycle means that the movements throughout locomotion or array of events or time period, where one foot of persons or any animals is in contact with the ground to when that same foot again contacts the ground, and that involves propulsion of the center of gravity in the direction of motion. Another name of this is strideGait [1]–[3].

Nowadays, transportable or wearable smart electronics devices are used frequently for different purposes and they are made with the various type of sensors. Though their demand is much high, these devices are developed swiftly, and the manufacturing companies are improving these devices according to the demand of the owner and also depend on the health condition of the regular people [4]. The gait Inertial sensor which is including gyroscope or accelerometer or both in it is progressively being attached in profitable transportable electronic devices. Some example of this are smartphones, they are very effective and high in cost but their performance is better and these inertial sensors are mainly used in the active research topic [5]. Gait inertial sensor data was made by Osaka University- Institute of Scientific and Industrial Research (OU-ISIR) [6]. They have proposed that they have the largest subject for an inertial dataset like 744. They have also mentioned some advantages for their dataset. Those are: (a) Walking down, and upslope and also on a flat ground-these three ground slope conditions are found, (b) A wide-ranging (2 to 78 years) ages, (c) Balanced gender ratio like 1, (d) Locations of the sensors are difference on subject's waist and, (e) Together angular velocity and acceleration data are taken by three inertial sensors like (including equally gyroscope and accelerometer) and a smartphone which contain an accelerometer in it [6], [7].

For doing this work we have explored supervised machine learning. In the process of feature extraction, a primary set of raw variables has been condensed to more manageable features that mean groups for processing. The grouping data are still completely and accurately describing the original data set [8], [9]. Segmentation means to divide or separate the dataset into many parts. It is very important in machine learning. In this paper, we have used the inertial sensor-based gait data for predicting the age and gender of people using supervised machine learning [10].

The necessary descriptions of the related works and literature review have been provided in section II. In section III, we have summarized the existing gait dataset relevant to our work. The methodology has been shown in section IV. Section V depicts the results we have experimentally obtained. Finally, we have concluded and discussed the future scope of our research in section VI and section VII respectively.

II. Literature Review

The technology for the analysis of gait, used by physiotherapists, includes treadmills, faceplates, insoles, instrumented shoes, and video tracking. These are used to measure the distribution of pressure under a persons feet and the movements of joints in different angles. Some effective methods are used for making the dataset, and it consists of the angle joints movement using low vision system. This dataset is composed or collected by two cameras, a treadmill, and numerous reflexive marks. For this dataset, they have used the deep and shallow architecture of machine learning algorithms. Deep machine learning is comprised of multi-levels of representation where the shallow has few levels of representation. Another team suggests that they have used motion recording sensor for collecting dataset of gait. There used a belt which includes MR sensors for collecting data from the subjects. Acceleration is recorded in three orthogonal directions, was attached with the subject's belt [5], [11].

Along with the previous, another group used some of the commercial smartwatches for gait-based biometrics. In that

978-1-7281-0788-2/19/1.00 ©2019 IEEE

case, there is a large number of prior works is based on that. Maximum of gait based biometric work are separated by the wearable sensor or machine vision based. This work is based on the real-time data of surveillance [12].

In this paper, we have used the inertial sensor data. There we have used the dataset for supervised learning. There are many available databases has been used in this specific area. Most of them have a low number of subjects like 58, 60, 36, 50. These databases do not include the age and gender of a person in it. Name of some datasets is Kobayashi et al. [2], Derawi et al. [13], Jenifer et al. [14] and Gafurov et al. [15], [16]. From these dataset Kobayashi et al. have the largest dataset based on smartphones. 40 out of 58 subjects are captured on different days. These datasets are clearly giving us a view about the biased of age distribution and gender ratio. For these types of dataset, some scholars use 3 types coefficient of Fourier transform. They are gait model-based methods, frequency domain analysis methods, and period detection-based methods. Period detection implies on the real-time application but frequency domain analysis does not have this. Here one method is gait model which is used for the gait pattern of homogeneity and equilibrium. For gait-based analysis, it needs expert knowledge of human gait for the machine and it has a lacking and that is it does not have the ability to deal with the signal-vibration [17].

III. Dataset Interpretation

The OU-ISIR Gait Database, Inertial Sensor Dataset has been used by our team for solving the problems on regression and classification. The database is quite well structured and is the largest inertial sensor-based gait database [18]. The database takes into account both accelerometer and gyroscope of three inertial measurement units and a smartphone around the waist of a subject. The dataset contains a total of 744 subjects (389 males and 355 females) with ages ranging from 2 to 78 years. There are two major subsets of the dataset namely: i) Automatic Extraction subset, and ii) Manual Extraction subset. The Automatic Extraction subset includes a wide range of about 744 subjects with just 2 activity labels, center-sensor level-walk sequence 0 and center-sensor level-walk sequence 1. On the other hand, the Manual Extraction subset includes four other partitions by the names; Android, IMUZ Center, IMUZ Left and IMUZ Right with 406, 494, 493 and 492 subjects respectively and each partition have 4 activity labels; slope-down, slope-up, level-walk 1 and level-walk [19], [20]. We have designed a graphical representation to depict a summary of this dataset which is shown in Figure 1.

IV. Methodology

We have begun with signal pre-processing of the raw signals to remove unwanted noise. Then we have segmented the entire dataset using the sliding window technique. After the segmentation process, we have constructed the feature vector. We have extracted statistical and energy based features from each segmented window. We have recorded the labels which correspond to each sample of the feature vector, which was

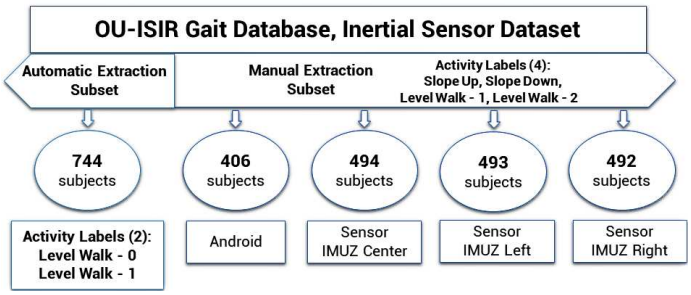


Fig. 1: Summary of the interpretation of OU-ISIR Gait Database, Inertial Sensor Dataset.

combined with the feature vector to form the train set and test set. Finally, we have trained our desired model by using the train set. For the evaluation of our model, we have adopted the N-fold cross-validation technique. In the following section, we have described our method. The proposed methodology has been summarized graphically in Figure 2.

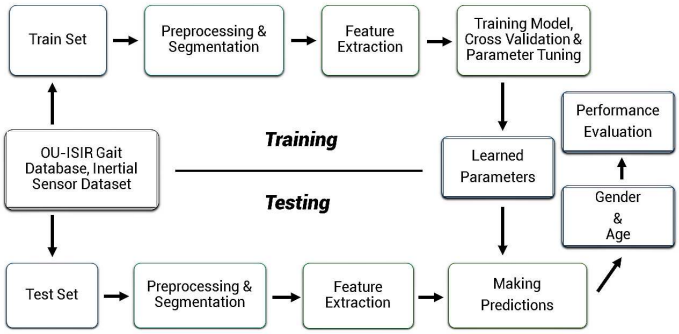


Fig. 2: Graphical illustration of the proposed methodology.

A. Preprocessing and Segmentation

For eliminating noise, we have used a low pass Butterworth filter. The operating frequency of accelerometer is 100Hz. But the human walking frequency is limited within 20Hz. Therefore, we have chosen 20Hz as the corner frequency of the filter.

In the dataset package, there are 1490 files with sensor values in x, y, and z-axes, provided as gait data for the automatic extraction subset of the given database. Each of these 1490 files is arranged in pairs where every pair belongs to a particular subject but they differ from one another in only the activity label. Of the pairs, the first file always has the activity label of level-walk sequence 0 and there, the last file always has the activity label of level-walk sequence 1, both of which belongs to a single subject or person whose age and gender have been separately recorded previously. Again, each file of the total 1490 files contains a number of samples. The total sample number in each file is not uniform and is found to be in the range from 300 to 1200. According to previous research, by using a 3s sliding window, the walking pattern can be illustrated with higher precision [21]. Therefore,

for segmentation, we have chosen a 50% overlapping sliding with a duration of 3s. Each window contains 300 samples, where 100Hz is the frequency at which the sensors operate in this setup. This method is then similarly applied for the Manual Extraction subset of the dataset. Here, every single file contains 6 columns of the dataset, in which three are for raw accelerometer data (ax, ay, and az) and the remaining three are for raw gyroscope data (gx, gy, and gz). Consequently, each file is then converted to a separately formed vector of ax, ay, az, gx, gy and gz values. In the meantime, the samples with the subject IDs are noted separately and are paired with the corresponding age and gender list that is provided.

B. Feature Extraction

After segmentation, we have extracted useful features that are supplied as the input of the predictor. Feature extraction is the most important part of the classification. Twelve types of time-domain features have been chosen for extraction from the segmented dataset. The features are mean, median, maximum, minimum, mean absolute deviation, standard error of mean, skewness, kurtosis, standard deviation, variance, root mean square, vector sum. Since our main goal is to identify the pattern of walking on different surfaces (e.g., plane, stair and incline), we have extracted features related to statistical and energy information of motion. From our research on similar works, we have found that these features are suitable for classification and regression problems from time-series inertial sensor-based data [22]. By mean feature, we can determine the average acceleration, which will vary from age to age. A young person can walk faster than an older person. Similarly, the median feature of acceleration also differs based on age. Biologically, the walking pattern of men and women of different ages are different. Therefore, for an accurate depiction of the walking pattern, we have derived more statistical features. In addition to extracting statistical features, we have also extracted a good number of energy related features. The rms feature is the typical root mean square value of the sensor values along the x, y or z-axis within a particular window frame of 300 samples. On the other hand, vs represents the vector sum of the sensor values along x, y and z-axes at each instant. The energy features provides information of the energy of motion which can be interpreted effectively to determine gender and age. For instance, the energy of walking of the male population above 20 years of age are generally high. In Table I we have mentioned all the features used for classification and regression.

In Table I the feature notations used are described below: max: maximum; min: minimum; mad: mean absolute deviation; skew: skewness; kurt: kurtosis; std: standard deviation; var: variance; rms: root mean squared; sem: standard error of mean (unbiased); vs: vector sum.

TABLE I

mean-ax	min-gy	kurt-az	vs-axyz
mean-ay	min-gz	kurt-gx	vs-gxyz
mean-az	mad-ax	kurt-gy	vs-mean-axyz
mean-gx	mad-ay	kurt-gz	vs-mean-gxyz
mean-gy	mad-az	std-ax	vs-median-axyz
mean-gz	mad-gx	std-ay	vs-median-gxyz
median-ax	mad-gy	std-az	vs-max-axyz
median-ay	mad-gz	std-gx	vs-max-gxyz
median-az	sem-ax	std-gy	vs-min-axyz
median-gx	sem-ay	std-gz	vs-min-gxyz
median-gy	sem-az	var-ax	vs-mad-axyz
median-gz	sem-gx	var-ay	vs-mad-gxyz
max-ax	sem-gy	var-az	vs-sem-axyz
max-ay	sem-gz	var-gx	vs-sem-gxyz
max-az	skew-ax	var-gy	vs-skew-axyz
max-gx	skew-ay	var-gz	vs-skew-gxyz
max-gy	skew-az	rms-ax	vs-kurt-axyz
max-gz	skew-gx	rms-ay	vs-kurt-gxyz
min-ax	skew-gy	rms-az	vs-std-axyz
min-ay	skew-gz	rms-gx	vs-std-gxyz
min-az	kurt-ax	rms-gy	vs-var-axyz
min-gx	kurt-ay	rms-gz	vs-var-gxyz

C. Model Training

For evaluation and training, we have split the whole dataset into three parts. Before preprocessing, we have segregated 15% data from the dataset, which have been considered as the test set. Remaining 85% data has been acknowledged as the train set. Due to the lack of enough data, we have to choose this ratio of data for separation. Then we have processed these two sets into two different thread. After pre-processing, we have obtained 7,785 samples for training and 1,374 samples for testing. As we have pre-processed testing data into a different thread, the possibility of test data leaking out to train data is zero.

For the training session, we have trained our model using the training set with 7,788 samples first. Then 15% of this training set has been split for 10-fold cross-validation process to verify whether the model is overfitting the training set. Finally, the previously held out test set with 1,371 samples, which the model has never seen before, has been used to determine how well the model generalizes on the test set data. In the following section, we have described the classifier and regressor model.

1) Classification: For classification purpose, i.e. to predict the labels 0 or 1, where 0 represents female and 1 represents male, we trained a number of well-known classifiers. From our results, we chose three with best performance namely k-Nearest Neighbors (kNN) [23], Support Vector Machine (SVM) [24], [25] and Random Forest (RnF) [26]–[28] classifiers. It should be mentioned that cross-validation accuracy and other performance metrics were quite low, initially, which was made satisfactory through various hyperparameter tuning.

2) Regression: For regression purpose, i.e. to estimate the age of the given subjects, we trained some well-known regressors out of which the following gave us good performance: k-Nearest Neighbors Regression (kNNR) [29], Support Vector Regression (SVR) [30], [31] and Decision Trees Regression (DTR) [32]. Just like classification problem, the cross-validation errors were initially quite high and tuning of individual hyperparameters was required.

V. Result and Analysis

After the completion of model training, the trained models are used to make predictions for both classification and regression problems. We have also shown the k-fold cross-validation results for 10 iterations over the training set to ensure that the model did not overfit the training set. We have also noted the particular hyperparameters of each model and the best test set accuracy obtained from each model. The results are shown in the following two subsections.

A. Performance Evaluation of Classifiers

The performance of the trained algorithms for classification of gender (male and female) is evaluated in this section. Table II shows the average cross validation accuracy and F1 score of each of the models (with chosen hyperparameters and respective confidence interval in braces) for 10 iterations. It also shows the test set accuracy. Figure 3 shows the confusion matrix of each model for gender classification. Figure 4 shows the decision boundary plotted by the classifier models on the 2D feature space-test set formed by using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique. This technique was used as we have a multidimensional feature space (88D) which is difficult to visualize. As a result, we represent the feature space as an equivalent 2D feature space using t-SNE method.

TABLE II

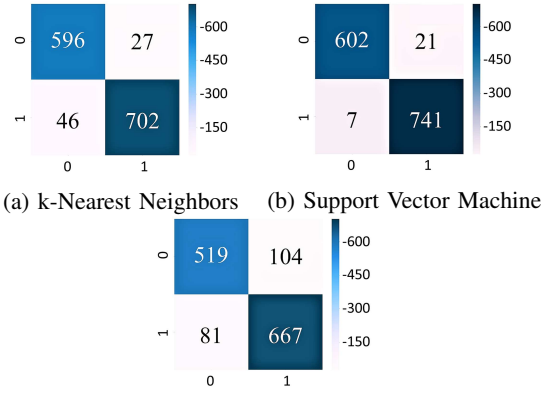
Classifier	Hyper parameter	Cross Validation Accuracy	Cross Validation F1 score	Test Set Accuracy
K-Nearest Neighbors	n_neighbor = 3	88.68% (1.66%)	89.06% (1.57%)	81.18%
Support Vector Machine	kernel = 'RBF'	92.66% (1.57%)	93.05% (1.43%)	84.76%
Random Forest	n_estimator = 25	86.55% (3.54%)	87.27% (3.15%)	73.67%

From Table II, it is evident that SVM outperforms both kNN and Random Forest classifiers in the task of gender classification, in terms of 10-fold cross validation accuracy and F1 score in addition to test set accuracy. The hyperparameters which were effectively tuned in order to obtain the best performance from the classifiers are manifested in the table. Confidence intervals for 10-fold cross validation accuracy and F1 score are mentioned in braces. In Figure 3, we have the confusion matrix of the classifiers. We have also demonstrated the decision boundary in Figure 4.

From Figure 4, we can confer that, k-Nearest Neighbors is able to make accurate decision boundaries to separate the two classes of samples. Support Vector Machine also generalizes well on the samples as well as on the outliers. On the contrary, Random Forest generalizes well on the samples but also captures the outliers which result in poor test set accuracy.

B. Performance Evaluation of Regressors

The supervised regression algorithms are evaluated in this section. Table 3 shows the average R^2 score along with mean squared error, mean absolute error and median absolute error of each of the models for the prediction of age. Figure 5 shows the estimations made by the regression models on the test set with 1D feature space formed by using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique as visualizing an 88D dataset is immensely difficult.



(a) k-Nearest Neighbors (b) Support Vector Machine

(c) Random Forest

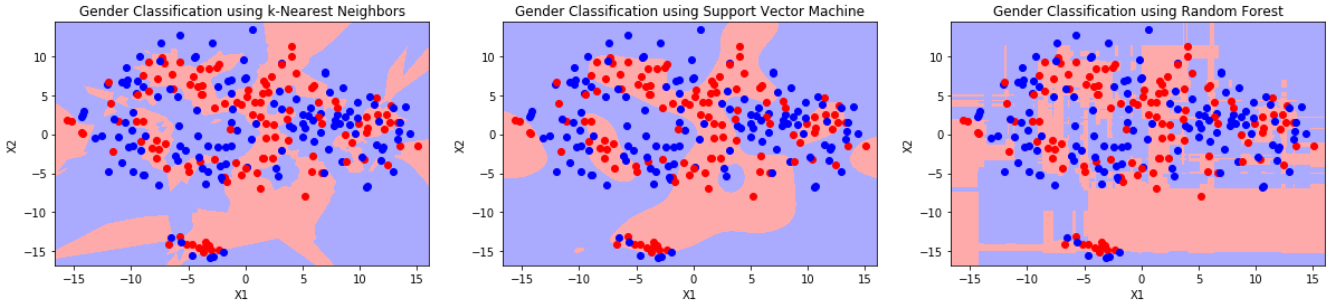
Fig. 3: Confusion Matrix for classification of gender (1 denotes male and 0 female)

TABLE III

Regression Model	Hyper parameter	Variance Score R^2	Mean Squared Error
K-Nearest Neighbors	n_neighbor = 3	0.54	0.46
Support Vector Machine	kernel = 'RBF'	0.55	0.45
Decision Trees	max_depth = 25	0.64	0.36

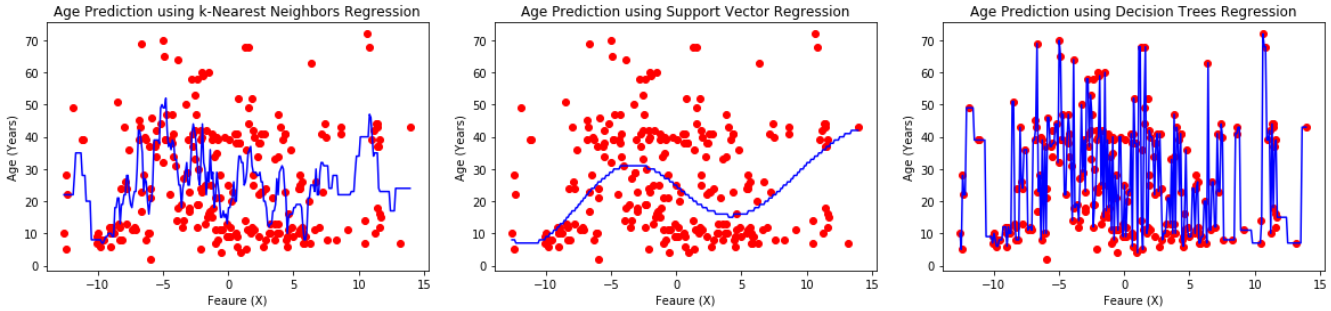
Table III represents the values of performance metrics for the models, kNNR, SVR and DTR. Out of the four metrics, variance score or R^2 is the most well-known metric for performance evaluation of a regressor. R^2 typically determines the correctness of the total number of predictions. It may have the values ranging from 0 to 1 where 1 represents 100% correct predictions and 0 represents 0% correct predictions.

So, it is quite clear from the table that DTR model is best suited for the task of predicting age in comparison to kNNR and SVR. The number of neighbors chosen to tune kNNR is 3. On the other hand, SVR with radial basis function kernel was chosen. Finally, the parameter maximum depth of 25 was chosen to tune DTR so that it gives the best outcome. It should also be noted that the other metric mean squared error is also important. Mean squared error is a converse of the variance score as it indicates the overall errors in estimates made by the regression model. In our case, the mean squared error is consistent with the variance score which also shows that DTR has the highest possibility of making accurate predictions of age. It should be mentioned here that the features chosen to be used for training and testing the model are the same which



(a) k-Nearest Neighbors (b) Support Vector Machine (c) Random Forest

Fig. 4: Illustration of decision boundary (blue dots denotes male and red dots female)



(a) k-Nearest Neighbors Regression (b) Support Vector Regression (c) Decision Tree Regression

Fig. 5: Illustration of model estimates on a 1D test set (red dots denote samples, blue line represent estimate from learned parameters)

were used for the classifiers mentioned the previous section. This brings up a future prospect of extracting new features that has higher correlation to age estimation and thus have more effective contribution in predicting age.

From Figure 5, we have found that k-Nearest Neighbors Regression makes some accurate predictions for lower feature values (X) but fails to predict the majority of the samples for higher feature values (X). Support Vector Regression makes a very simple non-linear fit to the sample data points. However, such type of model estimate is quite helpful to avoid overfitting and acquire good R^2 score both on the train and test sets. Finally, Decision Trees Regression makes a complex non-linear fit to the data. This provides higher R^2 score on the train and test sets and thereby, the lower rate of prediction errors. The difference between the predicted values and the actual values is the lowest in this case.

VI. Conclusion

In this paper, we have made an endeavour to determine the most optimum machine learning solutions to the problem of predicting gender and age from an inertial sensor-based gait dataset. We have experimentally analyzed the largest inertial sensor-based, OU-ISIR gait database for this purpose. In order to achieve our goal, we have proposed a pipeline of tasks which have been described throughout the paper. One of the key aspects of our paper is the use of time-domain based statistical features. We have successfully implemented these extracted features to train our chosen classification and regression models. We obtained some promising results which have been analyzed by time domain in section V. However, a common trend observed for all the classifiers is the higher accuracy for the training set compared to the test set. The distribution of the samples according to the age label seems rather random then structured which creates a barrier for the algorithms to predict the ages in an orderly manner.

VII. Future Work

The higher accuracy for the training set compared to the test set provides the sense of the problem of overfitting. This may be solved using regularization techniques like L1, L2 or Dropout Regularization, which constitutes an interesting research problem for future. In addition to that, Artificial Neural Network (ANN) with deep networks may be implemented with the anticipation of achieving better accuracy scores. Finally, more features can be extracted by extracting features in the frequency domain which will definitely aid in the training of the learning models. On the other hand, higher R^2 scores can be achieved when features having a higher correlation to the age label are used. We aim to extract more features in the near future which will have a higher correlation to the age label of the dataset.

REFERENCES

[1] Q. Riaz, A. Vögele, B. Krüger, and A. Weber, "One small step for a man: Estimation of gender, age and height from recordings of one step by a single inertial sensor," *Sensors*, vol. 15, no. 12, pp. 31 999–32 019, 2015.

[2] T. Kobayashi, K. Hasida, and N. Otsu, "Rotation invariant feature extraction from 3-d acceleration signals," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 3684–3687.

[3] T. P. Luu, K. Low, X. Qu, H. Lim, and K. Hoon, "An individual-specific gait pattern prediction model based on generalized regression neural networks," *Gait & posture*, vol. 39, no. 1, pp. 443–448, 2014.

[4] A. Muro-De-La-Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[5] S. J. Abbass and G. Abdulrahman, "Kinematic analysis of human gait cycle," *Al-Nahrain Journal for Engineering Sciences*, vol. 16, no. 2, pp. 208–222, 2013.

[6] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication," *Pattern Recognition*, vol. 47, no. 1, pp. 228–237, 2014.

[7] N. T. Trung, Y. Makihara, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi, "Phase registration in a gallery improving gait authentication," in *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–7.

[8] J. Brownlee, *Master Machine Learning Algorithms: discover how they work and implement them from scratch*. Jason Brownlee, 2016.

[9] L. S. Lincoln, S. J. M. Bamberg, E. Parsons, C. Salisbury, and J. Wheeler, "An elastomeric insole for 3-axis ground reaction force measurement," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2012, pp. 1512–1517.

[10] M. Sayed, "Biometric gait recognition based on machine learning algorithms," *Journal of Computer Science*, vol. 14, no. 7, pp. 1064–1073, 2018.

[11] M. O. Derawi, C. Nickel, P. Bours, and C. Busch, "Unobtrusive user-authentication on mobile phones using biometric gait recognition," in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2010, pp. 306–311.

[12] D. Gafurov, "Security analysis of impostor attempts with respect to gender in gait biometrics," in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2007, pp. 1–6.

[13] M. O. Derawi, P. Bours, and K. Holien, "Improved cycle detection for accelerometer based gait authentication," in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2010, pp. 312–317.

[14] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2010, pp. 1–7.

[15] D. Gafurov, E. Snekkenes, and P. Bours, "Spoof attacks on gait authentication system," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 491–502, 2007.

[16] D. Gafurov, "Performance and security analysis of gait-based user authentication," 2008.

[17] A. H. Johnston and G. M. Weiss, "Smartwatch-based biometric gait recognition," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2015, pp. 1–6.

[18] J. Lu and Y.-P. Tan, "Gait-based human age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 761–770, 2010.

[19] Y. Makihara, N. T. Trung, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi, "Phase registration of a single quasi-periodic signal using self dynamic time warping," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 667–678.

[20] N. T. Trung, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "Performance evaluation of gait recognition using the largest inertial sensor-based gait database," in *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 2012, pp. 360–366.

[21] A. D. Antar, M. Ahmed, M. S. Ishrak, and M. A. R. Ahad, "A comparative approach to classification of locomotion and transportation modes using smartphone sensor data," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2018, pp. 1497–1502.

[22] S. S. Saha, S. Rahman, M. J. Rasna, T. B. Zahid, A. M. Islam, and M. A. R. Ahad, "Feature extraction, performance analysis and system design using the du mobility dataset," *IEEE Access*, vol. 6, pp. 44 776–44 786, 2018.

[23] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 1. IEEE, 2010, pp. 91–94.

[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springe/r, 1998, pp. 137–142.

[25] D. Fradkin and I. Muchnik, "Support vector machines for classification," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 70, pp. 13–20, 2006.

[26] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. Apr, pp. 1063–1095, 2012.

[27] M. Denil, D. Matheson, and N. De Freitas, "Narrowing the gap: Random forests in theory and in practice," in *International conference on machine learning*, 2014, pp. 665–673.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] H. Dubey, "Efficient and accurate knn based classification and regression," *A Master Thesis Presented to the Center for Data Engineering, International Institute of Information Technology, Hyderbad-500*, vol. 32, 2013.

[30] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[31] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[32] M. Brijain, R. Patel, M. Kushik, and K. Rana, "A survey on decision tree algorithm for classification," 2014.