Gait recognition accuracy comparison using cycle-based segmentations with different classifiers on large accelerometer datasets

Arthur Gartner   
*Department of Cyber and Computer Sciences*   
*The Citadel*Charleston, United States  
AGartner@citadel.edu

*Abstract*—This paper reviews the accuracy attained from using a custom gait-cycle segmentation method with three different classifiers to match an individual to their recorded gait. Gait was measured using 3-axis accelerometer data from a mobile phone. The measurements were retrieved from a public database of 93 unique individuals walking at a steady pace between two end points. The cycle-based approach implemented a custom algorithm for cycle identification. The accuracy of the classifiers was tested using different thresholds for peak identification within the cycle-segmentation algorithm. The segmented data sets were then implemented in random forest, support vector machine, and multilayer perceptron classifiers with accuracy readings recorded and compared. The results of the experimentation indicate that random forest classifier provides the highest level of accuracy among the three classifiers when using cycle-segmented accelerometer data for gait identification.

Keywords—Gait, Accelerometer, Random Forest Classifier, Support Vector Machine, Multilayer perceptron, Classifier, Frequency, Cycle, Segmentation

# Introduction

The importance of securing mobile devices continues to rise as the capabilities and utility of these devices continues to increase. These devices are quickly replacing wallets and keys while holding even more personal information such as health and financial records. While biometric authentication in mobile devices has grown quickly with on-device fingerprint, iris, and face scanners these methods only attempt to authenticate a user when a request to unlock the device is made. While this is a secure biometric method for active entry into the device there currently exists no passive biometric security feature for mobile devices to authenticate a user while the device is locked and being carried. While locked the device is still capable of broadcasting unique identifiers over a multitude of wireless methods such as Bluetooth, Wi-Fi and Ultrawideband. This is increasingly important as the number and implementation of Internet of Things (IoT) devices increases. For convenience many IoT devices use the unique identifier broadcasted from a device to authenticate a user such as a connected home lock or a car with wireless key functionality such as a Tesla. Even financial transactions in the form of transit passes are now capable of being conducted with no further authentication required after initial setup. This increases the convenience for a user by not requiring any action from them. Authentication of the user is only required during setup of these IoT devices and is not required for continued passive operation. Overall security is reduced by assuming that a device broadcasting a unique identifier is in the possession of the original user who was originally authenticated and authorized during the setup and linking phase between the mobile device and IoT device. This produces a vulnerability where bad actors may be able to impersonate users simple by being in possession of their mobile device that continue to transmit the unique identifiers and allow passive access to previously linked IoT devices. Implementation of a passive biometric authentication method such as gait recognition would solve the potential vulnerabilities mentioned while maintaining the same level of convenience in passive interaction with IoT devices that users are used to. Previous published reports have already shown gait recognition accuracies of 99% using accelerometer data from a mobile device [1] providing credence to this approach.

The attempt of this paper is to provide results on accuracy readings on gait recognition using cycle-based segmentation with different classifiers. The data used for gait recognition will be in the form of x, y, z acceleration values provided by a mobile phone accelerometer with a sampling rate of about 100Hz. Data segmentation approaches taken for gait recognition using accelerometer outputs are primarily time or cycle based. A time-based segmentation approach breaks the accelerometer data into same time-length segments for. A cycle-based segmentation approach attempts to segment individual cycles within the data. This corresponds to an individual’s gait cycle which can be defined as starting when the right foot touches the ground and ends when the right foot touches the ground again. Cycle-based segmentation requires additional pre-processing steps, and this paper uses a custom approach to segmenting these cycles.

The classification techniques used within this paper are Random Forest, Support Vector Machine (Radial Basis Function kernel) and Multilayer Perceptron. These classifiers were chosen based on the results found in Implementation of machine learning algorithms for gait recognition [2] and Benchmarking the Performance of SVMs and HMMs for Accelerometer-Based Biometric Gait Recognition [3]. The results of the paper show a clear advantage for Random Forest and SVM classifiers with large accelerometer datasets when segmented by cycle lengths.

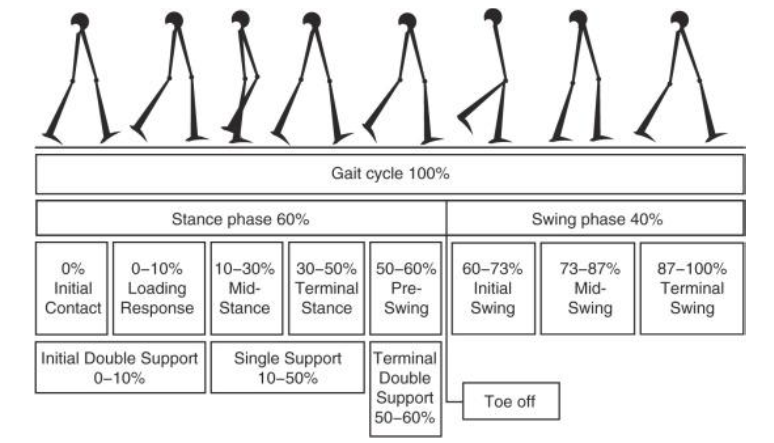
While many papers focusing on using machine learning techniques for gait recognition exist there are far fewer papers using cycle segmentation techniques with a larger 40+ person gait database. Most similar is the approach taken by Nickel and Busch [4] comparing cycle segmentation to time segmentation. This paper is different however due to the use of different classifiers (Random Forest, Support Vector Machine, and Multilayer Perceptron) along with a new cycle segmentation method.

The key contribution of this paper will be the process of segmenting the gait cycles as well as the accuracy data produced with such a large dataset. The dataset used for this paper is the Human Gait Database from Human Gait Database for Normal Walk Collected by Smart Phone Accelerometer [5]. It is important to note that the data in this dataset consists of 93 individuals of different ages and sizes, walking between two points at a constant, leisurely speed. Clothes, floor material and state of being (drowsy, invigorated, etc.) were not considered and could possibly affect gait measurements but this goes beyond the scope of this paper. Additionally, although two trials existed for most individuals as well as two mobile phone accelerometer values being recorded that were placed on different spots of the individual, only the accelerometer data from the first trial and the left waist placed mobile device was used.

Chart, bar chart, histogram

Description automatically generatedThe results of the experiment show an advantage of the Random Forest classifier over the two others in terms of accuracy for different thresholds of peak value identification within the segmentation algorithm. Random Forest Classifier displayed an average accuracy of 79.03% and 76.80% accuracy over datasets ranging from 5 to 50 individuals using a train/test ratio of 80/20. Support Vector Machine showed a slightly lower accuracy of 78.50% and 70.25% with Multilayer Perceptron coming in last at 72.30% and 66.51% accuracy. These results show a favor towards the Random Forest Classifier, but it is possible that Support Vector Machines could produce a high accuracy with modification of parameters.

# Related Works

Machine Learning based Human Gait Segmentation with Wearable Sensor Platform [6] presents the basis of the cyclic nature of gait. This paper explains the different phases of gait starting at initial contact to the terminal swing and can be seen in Figure 1. This paper does not attempt a random forest classifier and uses a different gait segmentation method than within this paper.

Using this cyclic approach accelerometer data can be segmented by cycle length and an individual can be linked to a cycle segment rather than a time segment. This allows the cycles to reinforce each other during training and allows for a shorter capture period.

Gait Identification Using Accelerometer on Mobile Phone [7] also conducts a similar approach as in this paper. Accelerometer data from a mobile device is used record an individual’s gait. Like this paper the authors chose to use only the z-axis value within the accelerometer reading values. The reasoning behind this choice is that the z-axis signal will have a stronger reading due to ground reaction and inertial force. Both these forces make the z-axis the most important axis for recording gait if a 3-dimensional vector incorporating all x, y and z axes isn’t used. The authors for this paper state that the intention of only using the z-axis values is to identify the true peaks which are the peaks with the highest magnitude generated from strong signal changes on the z-axis. For this research however the dataset used by the authors was much smaller with a total of 11 individuals and the remote forest classification was not used.

# Dataset Analyzation

Chart, bar chart, histogram

Description automatically generatedThe dataset used contains data for 93 individuals from various age and weight groups (Figure 2 and Figure 3).

The variance of age and weight within the dataset is important in verifying the accuracy of the proposed gait recognition methods and models among individuals of different biological characteristics. Human gait is influenced by weight and age so having a dataset consisting of a broad range of ages and weights can ensure that the gait recognition model is applicable to a broad range of individuals and produces an accuracy reading that is more applicable to real-world scenario versus perfect lab-controlled environments.

This dataset includes a multitude of sensor data for individuals walking between two points. For each session an individual carried two mobile devices. One was located on the right thigh and the other was located on the left side of the waist. During analysis of this data, it was found that although most individuals had two recording sessions there were several individuals who only participated in a single session as seen in Figure 4. To reduce the risk of overfitting only session 1 was used in this paper since that session included the most individuals. Additionally, not every individual has correctly linked datasets for the right thigh mobile device data in either sessions 1 or 2. Because of this missing data the left waist position dataset was only used for this paper.

Chart, bar chart

Description automatically generated

Despite the depth of sensor data available within the public dataset, only the accelerometer x, y, z values are used for analysis within this paper. The additional sensor information would be beneficial for future work and would perhaps help in increasing overall accuracy of the findings within this paper.

Finally, although 93 individuals were included in the dataset, this paper only evaluated the gait for the first 50.

# Proposed Methods

This paper will follow the process model shown in Figure 5.

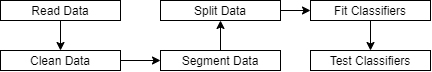
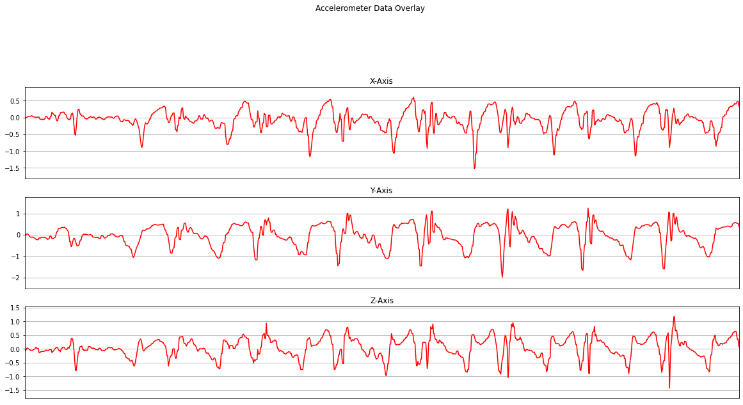
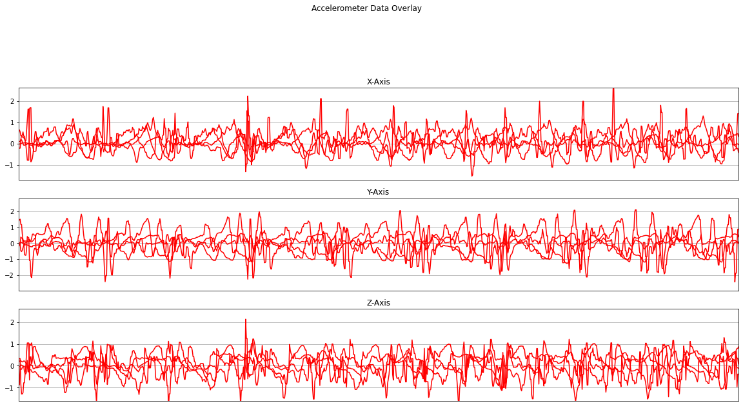


Figure 5: Process model for data

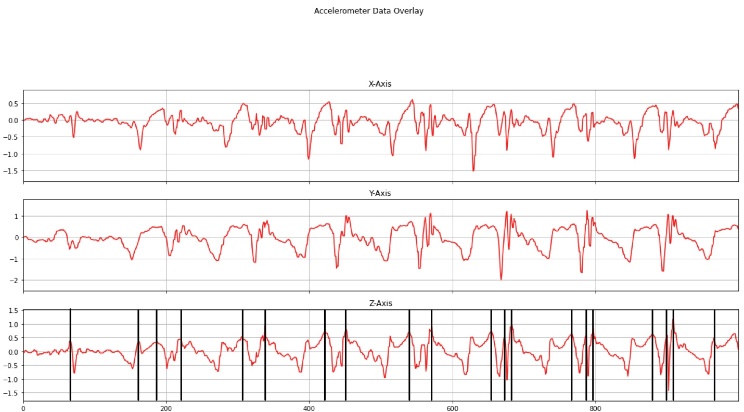
The first step in the process will require reading the data from .csv files and converting it into a data frame within Python. During this import process any individuals with missing linked datasets will be skipped. This is to ensure there are no null values or references within the imported dataset. Next the data will be segmented based on the custom cycle-based segmentation algorithm mentioned within this paper. For training and testing purposes the newly segmented datasets will be split into training and test data. This is necessary to fit/train each classifier and produce accuracy readings on newly seen datapoints. After completion of the process the accuracy reading will be totaled, and comparisons will be made between accuracy readings of the different classifiers and cycle-segmentation parameters.

## Reading the dataset

Using the public dataset each the x, y, z accelerometer values will be taken from each user’s first session from the mobile device located on their left waist. Since the sampling rate of the data appears to be around 100Hz for each user 1000 consecutive accelerometer readings will be taken starting from initial collection. This will account for about 10 seconds of gait accelerometer data for each user. Each 1000 sample sequence will then be plotted (sample seen in Figure 6) and tied to the user id of the individual they were captured from. An overlain view of multiple user’s accelerometer readings can be seen in Figure 7.

To properly segment the cycles an analysis of the axes must be completed. For this paper an analysis of only the z-axis will be conducted to determine cycle segments. The decision for this is based on the cyclic nature that the z-axis readings produce based on the individuals’ movements while walking. While other axes could be used for additional segmentation analysis the z-axis appears to be the most representational due to the upward movement of an individual’s legs as well as the impact force when an individual’s foot touches the ground.

## Segmenting the data points

Once all accelerometer values for the designated range of rows have been added to a data frame the process for gait cycle segmentation will begin. The process takes in two different parameters, a time analysis size, and a threshold of peak percentage value. The time analysis size will indicate the number of consecutive sequences to analyze as an entire unit and will be set to 300. This is necessary as the program requires a starting point to begin identifying the peaks within the dataset. The time analysis size also helps ensure that accelerometer values from multiple users are not categorized as just belonging to a single user.

The algorithm will take the time analysis size and begin review the range of values from the start of the dataset to the time analysis size. In this subset of data, the highest peak will be found. This process is completed by simply finding the maximum z-axis value recorded within the subset. Once the maximum peak z-axis value is recorded a threshold is calculated by multiplying the peak value with the threshold percentage parameter value. This newly calculated threshold value will now be used to determine other peaks within the subset. The threshold percentages used for this paper will be .04 and .08 to view accuracy differences as sensitivity for peak identification changes. After calculation of the threshold the algorithm will start at the beginning and iterate through all z values. If the z-value is above 0 then a check will be made to determine whether the point is a peak. This is done checking if it is greater than the local maximum and if it is above the previously calculated threshold. If the new value meets the requirements for a new local maximum, then the value is recorded and the next z value is evaluated. Once the z value is equal to or less than 0 then the range of z values between the start of the iteration and the peak will be added as a single new cycle-sequence. The start point for the next sequence will move to the index of the current local maximum which will then be reset to 0. The algorithm will continue iterating through 0 values until a new maximum is found. At this point the values from the previous peak to the current peak will be added as another cycle sequence. At the end of the subset an evaluation of the linked user id’s is conducted. Since the entire dataset is just a list of all the data it is possible to have minor overlap between user’s gait cycles. This occurs when a new sequence is initiated, and the start index is the index of the current peak. However, the potential error added is inconsequential if the time analysis parameter is of sufficient size. To account for improper labeling of overlapping gait cycles only the user id that shows up most often in the subset is used to link to the gait-cycle sequence. After linking a user id to the current segmented sequence, the next sequence is analyzed starting at the current index of the last known peak. An example of a segmented subset is shown in Figure 8 and 9.

## Converting the data

After creating of all cycle-segments the list data is required to be made into an array. Before an array can be made from this data the segments must be padded so that they are all the same length. The algorithm then identifies the longest segment cycle recorded and uses the length of the longest segment to pad the other segments with 0 values to achieve the same length. Once the lists have been padded successfully both the segments and the user lists are turned into arrays. The array produced by the segmented cycles list is a 3-D array following the format (number of segments, number of axes, length of segment). The number of segments value comprises of all segments created among all users. The index in this position matches the index in the 1-dimensional user array and allows for a link between the segment and the correct user-id. The number of axes value aligns with the x, y, z coordinates so this value is always 3. For the length of segment value this will be the number of the greatest segment length. Due to the padding requirement every segment’s x, y, z arrays will be of the same length as the longest segments x, y, z array.

After creation of the segmented 3-dimensional arrays the data is split into training and testing data. The ratio of train/test data used for this paper was 80/20. Both the test and train data are then needed to be converted into 2-dimensional arrays for use with the classifiers. To produce the 2-dimensional array the following format is used:

*(number of segments, (number of axes) \* (length of segment)*

This format produces a 2-dimensional array that can be used with the classifiers.

## Using classifiers

The classifiers used were Random Forest, Support Vector Machine (SVM) and Multilayer Perceptron. The random forest classifier was created with a value of 100 for the number of trees in the forest. The SVM was created with a value of C being 15 with a Radial Basis Function (rbf) kernel with a scaled gamma value. Although the higher value of C indicates low bias and high variance with a smaller margin hyperplane the test results did not appear to show any major effect. The multilayer perceptron was created with a single hidden layer with 100 neurons. Accuracy data was taken for each classifier and a confusion matrix was printed.

# Experimental Results

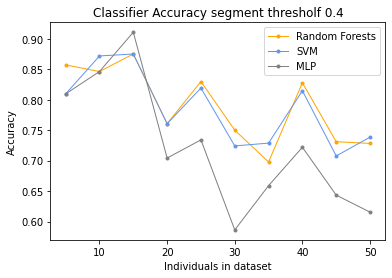
Due to the large size of the dataset only the first 50 individuals with associated datasets were used for this experimentation. For this experiment the accuracy was recorded from each classifier starting with 5 individuals and going to 50, with increments of 5 individuals. The segmentation algorithm was adjusted to identify any performance gains achieved by adjusting the threshold of which peaks were identified. The results of the experimentation are seen in Table 1 and Table 2.



Table : Accuracy results from different classifiers with varying individual enrollments with a cycle segmentation threshold of 0.4.



Table : Accuracy results from different classifiers with varying individual enrollments with a cycle segmentation threshold of 0.8.

Using the data from Table 1 and Table 2 associated scatterplots was created to visualize the changes over time as the number of individuals with gait measurements increased. These scatterplots are shown in Figures 10 and11.

As shown in Figures 10 and 11 accuracy of the classifiers decreased as the number of classification classes increases for both 0.4 and 0.8 peak threshold values. This is typical behavior as the algorithm has additional labels that increase the chances of a misclassification. Interesting behavior is shown between the random forest and SVM classifiers. For a 0.4 peak threshold these classifiers appear to follow the same trendline with no clear identification of the most accurate one. Taking the mean of accuracy over the collected values for each data point shows a slight advantage for the random forest classifier as seen in Table 3.

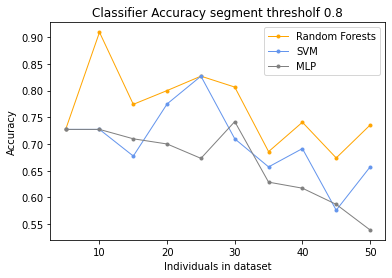
Figure 10: Scatterplot of classifier accuracy with cycle-segment peak threshold of 0.4

Figure 11: Scatterplot of classifier accuracy with cycle-segment peak threshold of 0.8

For a segmentation threshold of 0.8 random forest is clearly shown as being more accurate within Figure 11. The overall accuracy for random forest was 76.80% for this scenario. While lower than the accuracy achieved with a lower segmentation threshold, the lead in accuracy percentage when compared to the other classifiers is much greater with SVM and MLP both being several percentages lower.

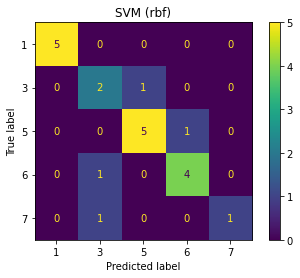


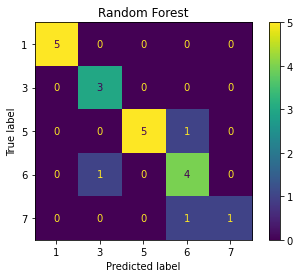
Table : Accuracy average over range of collected data points with segmentation threshold of 0.4

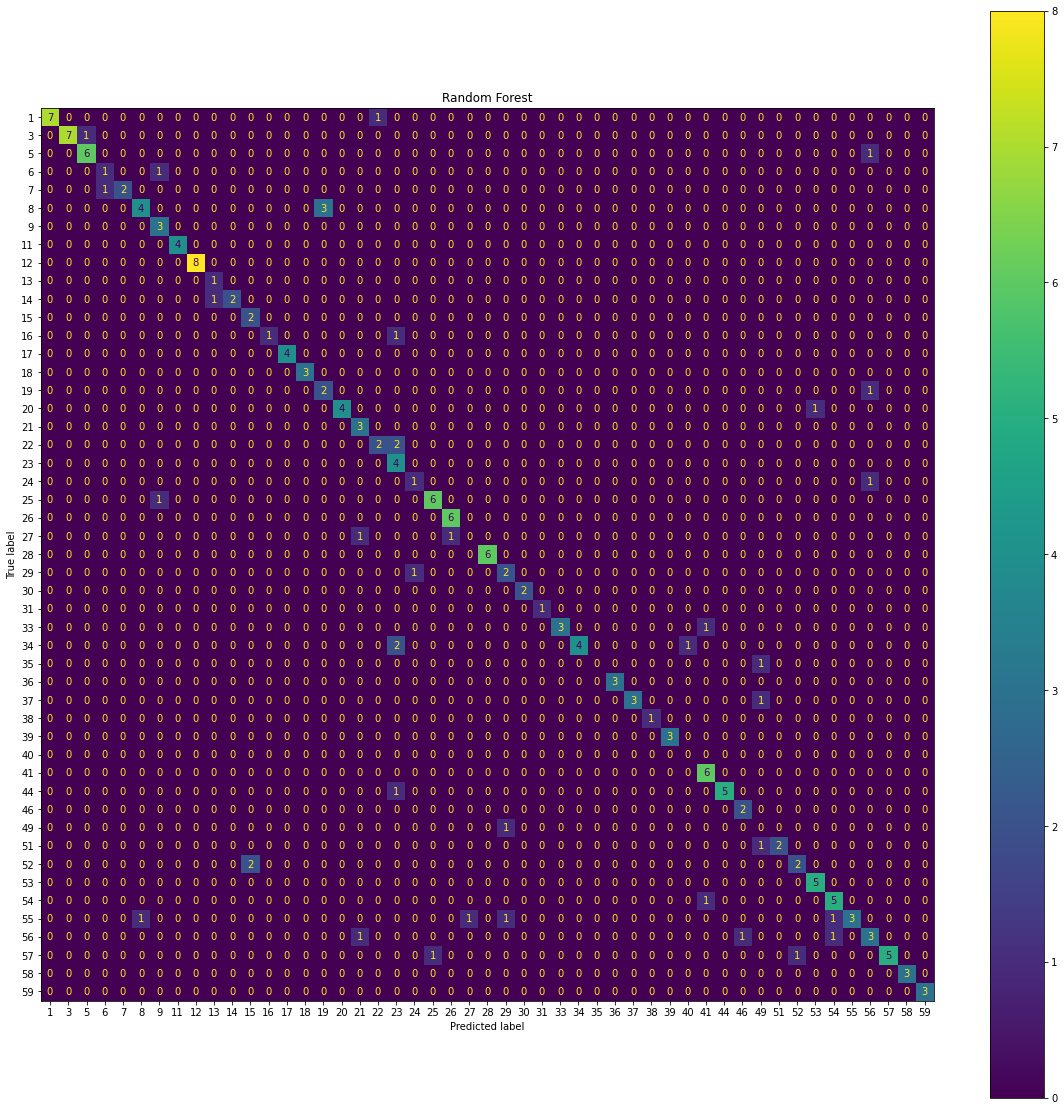


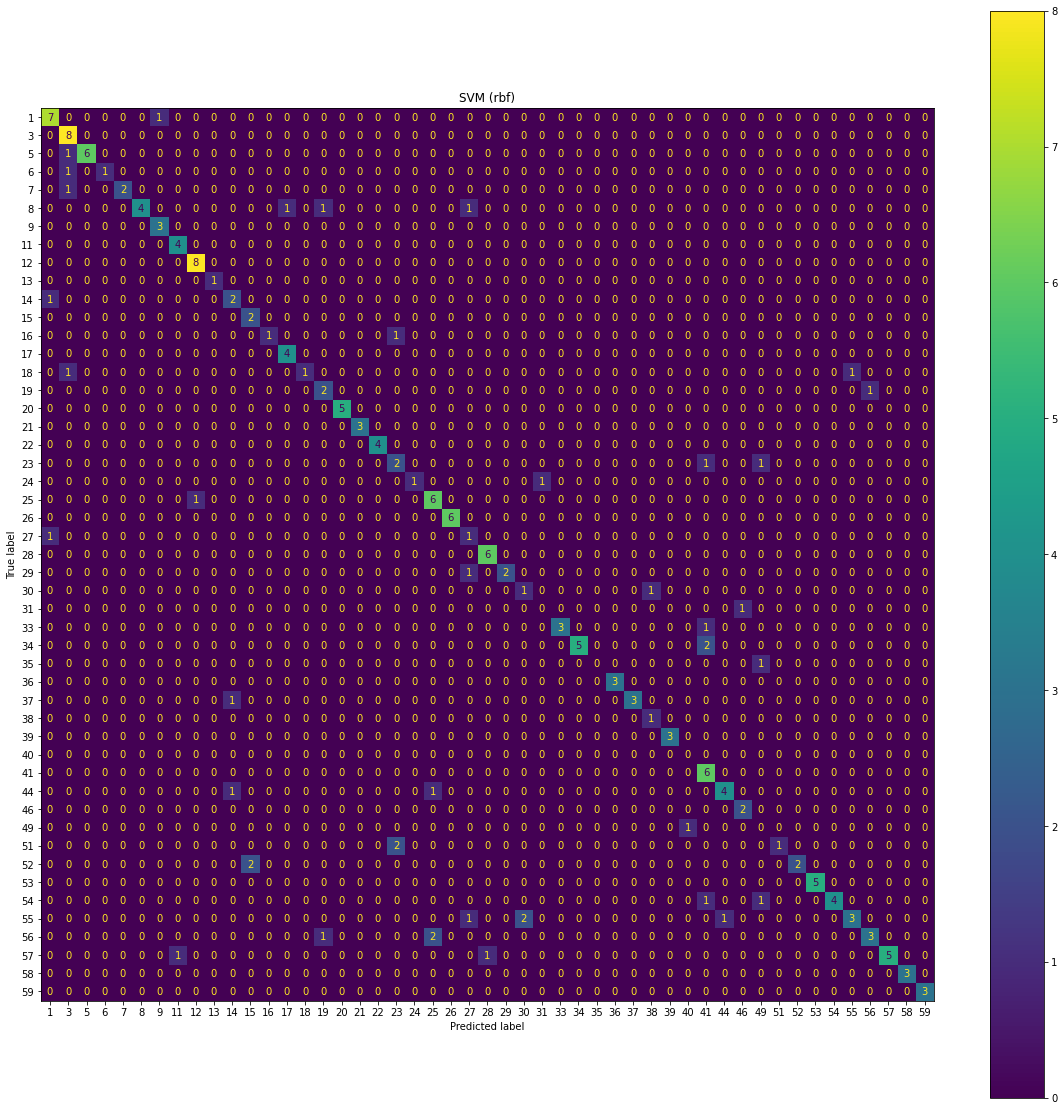
Table : Accuracy average over range of collected data points with segmentation threshold of 0.8

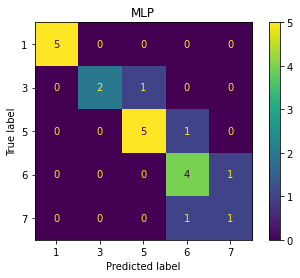
As evident from Table 3 random forest classifier had the highest accuracy beating SVM by 0.53% and MLP by 6.73% when a segmentation threshold of 0.4 was used. Using a segmentation threshold of 0.8 produced the average accuracies shown in Table 4 giving random forest classifier a lower average accuracy but a higher lead beating SVM by 6.55% and MLP by 10.29%.

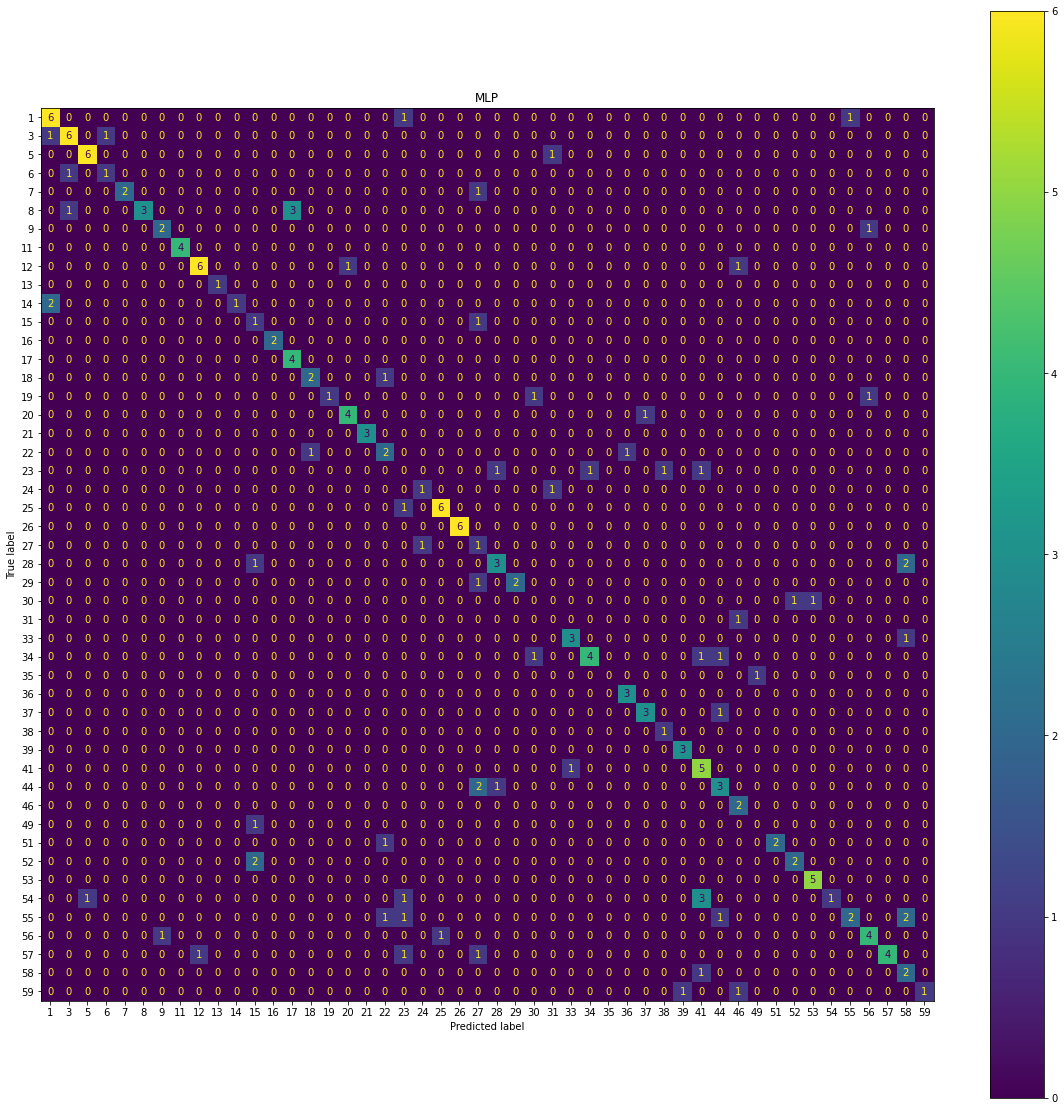
Another visual of performance can be produced by using a confusion matrix as seen in Figures 12 to 17. The confusion matrix was generated with an input of 5 individuals and an additional one was created for 50 individuals for each classifier.









 As evident in the confusion matrices the random forest and SVM classifiers also share many of the same misclassifications when a segmentation threshold of 0.4 is used. This is interesting because both classifications follow different methods for class identification. The MLP matrix shows much more erratic behavior with misclassification occurring across the board.

In comparison, the confusion matrix for the three classifiers with a total of 50 individuals are seen in Figures 18 to 20.

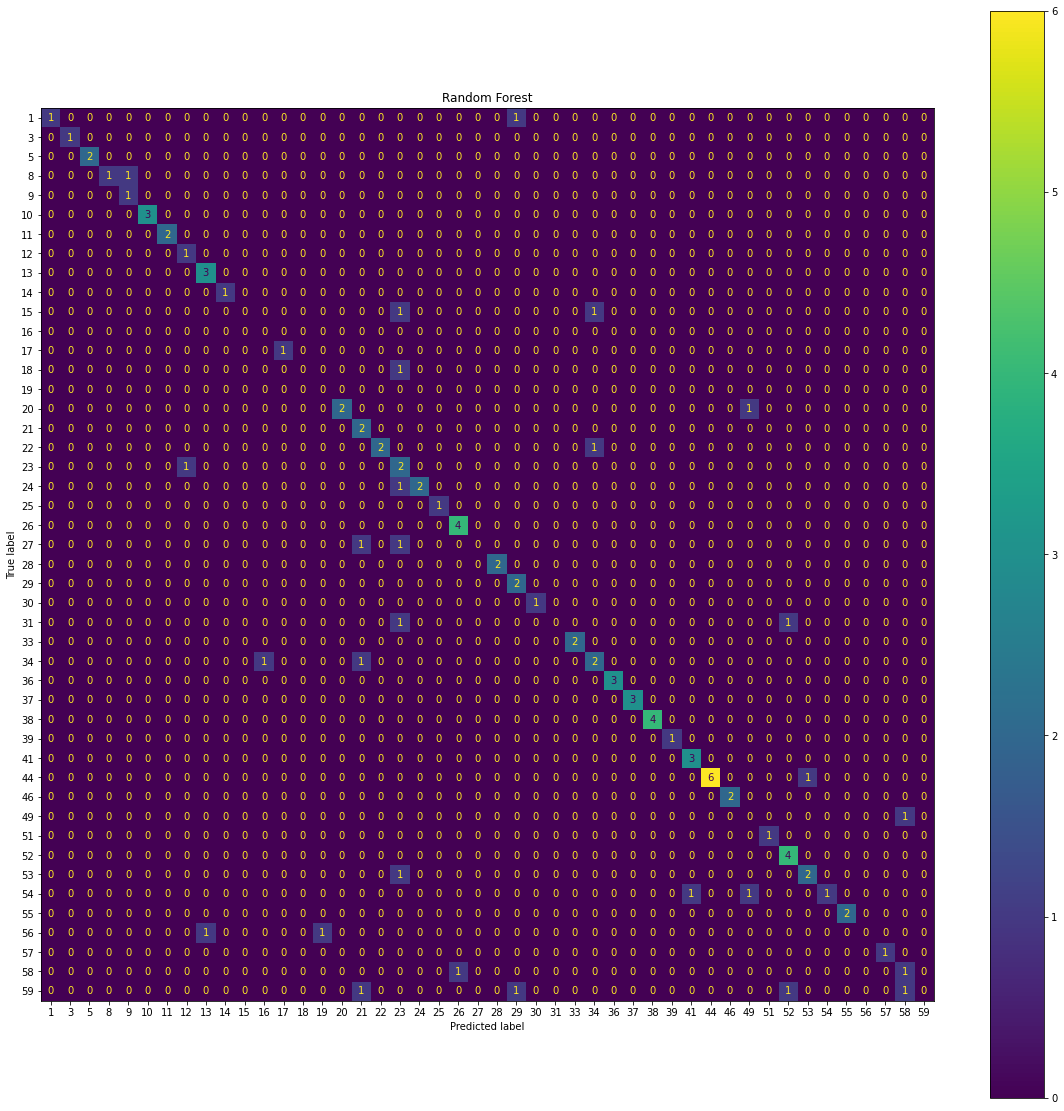
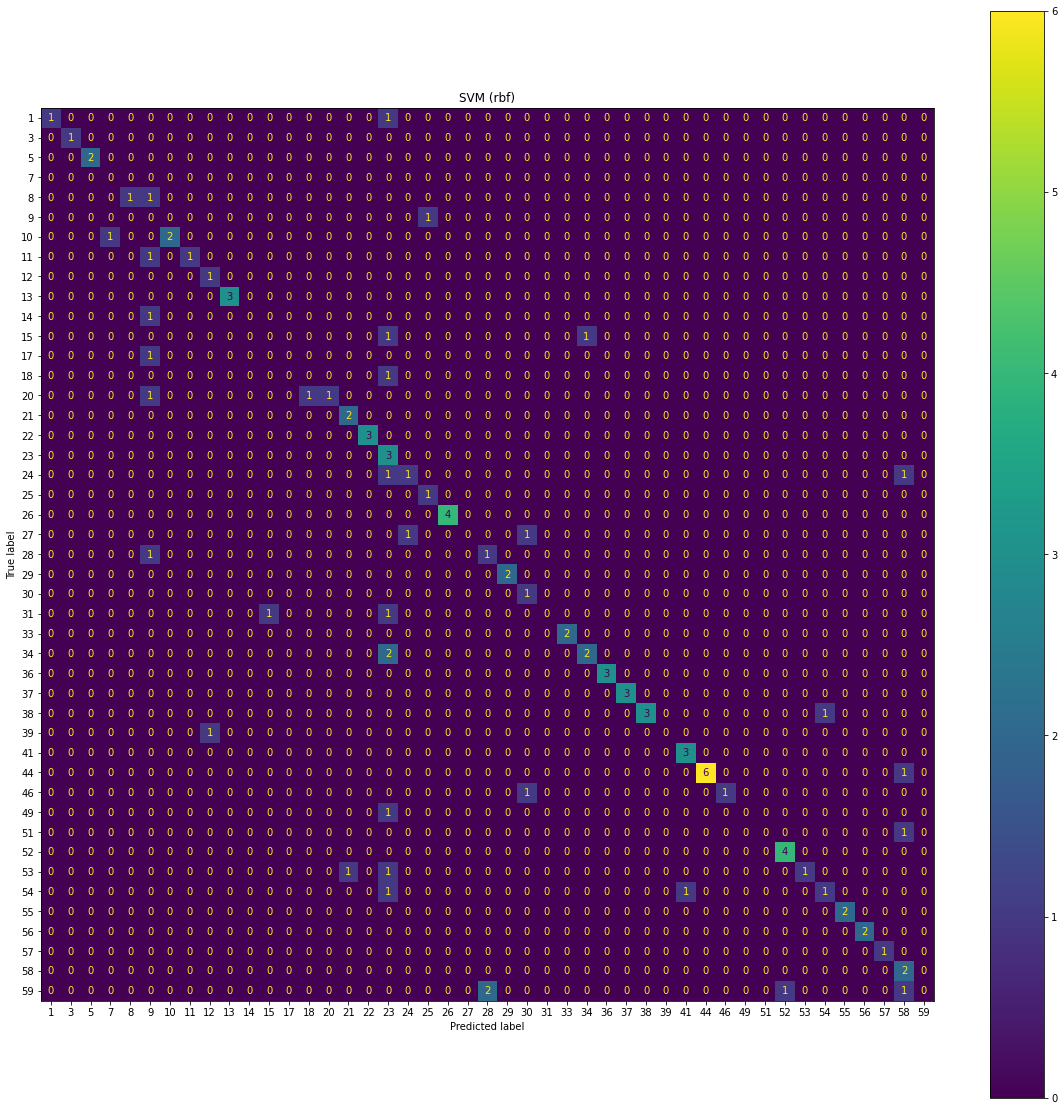
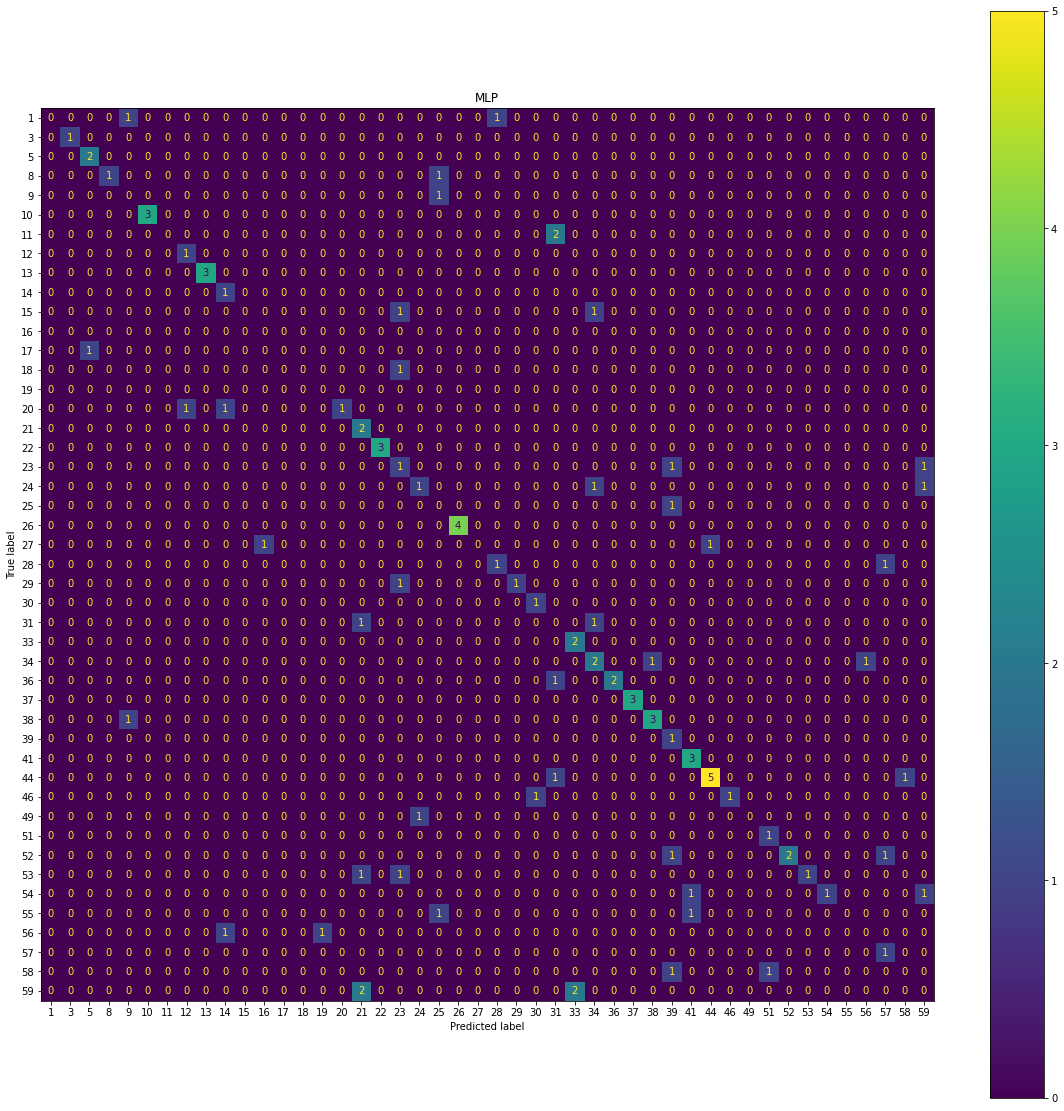


Figure 18: Random Forest classifier confusion matrix with 50 individuals (segmentation threshold = 0.8).



Reviewing the confusion matrices for the random forest and SVM classifier show that while they both made many of the same misclassifications the random forest was less sporadic than SVM, leading to the overall higher accuracy for random forest.

From these outputs the segmentation threshold can be seen to affect classifier accuracy. As the threshold increases, less segments are created. While cycles may be missed there is higher certainty that a segment contains at least one full gait cycle. Lower thresholds allow for sub gait cycles to be potentially treated as a full segment representing a full gait cycle. While this can be problematics it is interesting to see the higher accuracy readings when gait sub cycles are allowed to be segmented.

# Summary/Conclusion

The results of this paper indicate an advantage of the random forest classifier over a Support Vector Machine or Multilayer Perceptron classifier when attempting to identify individuals based on their gait after being segmented using gait cycles. Both segment thresholds of 0.4 and 0.8 show random forest classifier as being the most accurate classifier but the total accuracy and accuracy lead when compared to the other classifiers change when the threshold value is altered.

This paper explains a unique method of identifying gait cycles within the axis readings of a mobile device’s accelerometer. By using a parameter to determine a threshold for peak identification users can customize the sensitivity of the peak identification algorithm to identify better values based on their own dataset. Additionally, users may choose the time range of which to search for gait cycle segments. Variances of both these values may result in even higher accuracy readings and would be a potential focus for future work.

As the use of mobile devices for passive user identification increases the justification for passive biometric methods becomes stronger. Gait identification using onboard mobile device sensors is one viable method for passive biometric identification with gait-cycle segmentation methods in conjunction with random forest classifiers.

##### References

1. F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, and M. Savvides, “Gait-ID on the move: Pace Independent Human Identification using cell phone accelerometer dynamics,” *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012.
2. A. Kececi, A. Yildirak, K. Ozyazici, G. Ayluctarhan, O. Agbulut, and I. Zincir, “Implementation of machine learning algorithms for Gait Recognition,” *Engineering Science and Technology, an International Journal*, vol. 23, no. 4, pp. 931–937, 2020.
3. C. Nickel, H. Brandt, and C. Busch, “Benchmarking the performance of svms and hmms for accelerometer-based biometric gait recognition,” *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2011.
4. C. Nickel and C. Busch, “Does a cycle-based segmentation improve accelerometer-based biometric gait recognition?,” *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012.
5. A. Vajdi, M. Zaghian, S. Farahmand, E. Rastegari, K. Maroofi, S. Jia, M. Pomplun, N. Haspel, and A. Bayat, “Human Gait Database for Normal Walk Collected by Smart Phone Accelerometer,” Feb. 2020.
6. S. Potluri, A. B. Chandran, C. Diedrich, and L. Schega, “Machine learning based human gait segmentation with Wearable Sensor Platform,” *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.
7. H. M. Thang, V. Q. Viet, N. Dinh Thuc, and D. Choi, “Gait identification using accelerometer on mobile phone,” *2012 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2012.