Edge Hill University | Department of Computer Science

Edge Hill University | Department of Computer Science

CIS4519 Data Mining and Visualization

Coursework Two (CW2)

Exploring Data Visualization Techniques and Data Mining Solutions for Human Activity Recognition Using Wearable Sensors

Student Name: Sunday Idika
ID:****

Email: **@edgehill.ac.uk**

Table of Contents:

1. Introduction:

The effectiveness of visualizing data to uncover insights is vital in modern data analysis. This report focuses on using the Wireless Independent System Detection and Motion (WISDM) datasets (Gary, 2019) to find hidden patterns in motion-related data.

The WISDM datasets contain motion data collected from smartphones and smartwatches, covering various human activities. The raw data is visualized to gain useful insights, helping decision-making and hypothesis formation. Additionally, by examining different machine learning algorithms, inspired by previous research, the study seeks to understand which methods are most effective in accurately identifying activities based on sensor data.

The selection of the WISDM datasets stems from their richness in motion-based biometrics data, captured through smartphones and smartwatches. This dataset contains different activities, ranging from hand-oriented-eating to non-hand-oriented and hand-oriented-general facilitating a comprehensive exploration of human movement patterns. The rationale behind utilizing visualizations lies in their capacity to distill complex datasets into visually interpretable forms, enabling researchers to understand relationships and trends.

In the study conducted by Weiss, Yoneda, and Hayajneh (2019), different machine learning algorithms were used for modeling, Random Forest performed very well due to its renowned interpretability and efficiency in classification tasks, which were useful in their motion-based biometrics research. Similarly, Zhou et al. (2022) recommended Logistic Regression as a baseline model for human activity recognition, underscoring its relevance and utility in wearable sensor data analysis. Drawing insight from these studies, an acknowledgment to utilize the versatility and significance of the two machine learning approaches in addressing research objectives, thereby informing our methodological choices.

2. Data Visualization Tools and Analysis

Data visualization involves using different methods and tools to represent complex data visually, making it easier to understand and communicate insights (Shakeel et al., 2022).

Data visualization involves the application of various tools to represent complex data visually, aiding in understanding and communicating insights. In this section, the researchers explore how different visualization tools are applied to the Wireless Independent System Detection and Motion (WISDM) datasets to uncover patterns and relationships within the motion-related data.

Pandas:
Pandas serves as a crucial tool for data manipulation and analysis in the researchers' analysis. It efficiently organizes and preprocesses the WISDM datasets, leveraging its DataFrame structure for tasks such as data cleaning, transformation, and aggregation, facilitating the preparation of the dataset for visualization and analysis.

Matplotlib.pyplot:

Matplotlib.pyplot emerges as a fundamental tool for generating various types of plots and visualizations. It allows the researchers to create visual plots illustrating activity distributions, trends, and relationships within the WISDM datasets. With customizable options for colors, annotations, and plot types, Matplotlib.pyplot enables effective communication of insights derived from the data.

Seaborn:
Seaborn complements Matplotlib.pyplot by offering advanced statistical visualization functions. The researchers utilize Seaborn to create visually appealing and informative plots exploring complex patterns and relationships within the WISDM datasets. Its capabilities for visualizing distributions, correlations, and trends enhance understanding of the underlying data structure.

NumPy:
NumPy plays a pivotal role in numerical computations and data analysis. It supports efficient array operations and mathematical functions, aiding in calculations and transformations necessary for data visualization and analysis on large datasets.

Scikit-learn:
Scikit-learn, a versatile machine learning library, includes tools for data preprocessing, model training, and evaluation. While primarily used for machine learning tasks, it also offers functionalities for data visualization. The researchers leverage Scikit-learn to visualize classification results, such as confusion matrices, assessing the performance of machine learning models trained on the WISDM datasets.

By harnessing these visualization tools, the researchers gain deeper insights into the motion-related data captured in the WISDM datasets. Through visual exploration and analysis, they uncover patterns, trends, and relationships that inform understanding of human activities and behavior.

According to Kosara (2016), different visualization techniques serve specific purposes, in visualizing the WISDM dataset, the following plots were used:
Bar plots: Useful for showing activity distributions and frequency counts.
Box plots: Useful for understanding the data distributions, identifying outliers, showing central tendency, and the concentration of the data.
Scatter plots: Useful for visualizing the relationships between variables, helping identify correlations, clusters, or trends within the dataset.
Heatmaps: Useful for plotting confusion matrices of classification models and evaluating the model performance by comparing predicted and actual activity labels.
PCA plots: Useful for visualizing high-dimensional data in reduced-dimensional space, providing insights into data variance, and facilitating dimensionality reduction.
Confusion Matrix: Useful for classification model accuracy, confusion matrices enable the assessment of model performance strengths and weaknesses, thereby aiding in the interpretation of classification results.
Time Series Plots: These plots represent data collected over time intervals, revealing temporal trends and patterns in activity data recorded by sensors, including fluctuations, periodicities, and anomalies.

3. Implementation and Code

The provided code snippet (See attached video and code link submission) offers a structured framework including data preprocessing, model training, evaluation, and visualization, facilitating a comprehensive analysis of classification model performance. Initially, a careful sampling process selects 5% of the dataset, ensuring a representative subset for analysis. Subsequent data filtering focuses exclusively on walking and jogging activities from phone accelerometer data, enhancing the dataset's relevance by excluding extraneous data points. Feature selection follows, with accelerometer data (X, Y, Z axes) chosen as features and activity labels extracted, establishing the groundwork for model training. The dataset is then partitioned into training and testing sets to enable robust performance evaluation, ensuring models are trained on one subset and tested on another to mitigate overfitting risks (Zheng, Liu and Ge, 2022).

Moving forward, the code embarks on model training, starting with the instantiation and training of a Random Forest classifier, renowned for its ensemble learning capabilities (Jannat et al., 2023). Model evaluation ensues, leveraging the testing data to compute classification reports and accuracy scores, offering insights into the model's performance across various activity categories. Additionally, confusion matrices are generated and visually represented, providing a detailed breakdown of classification errors and model strengths, crucial for fine-tuning model parameters and improving overall performance (Valero-Carreras, Alcaraz and Landete, 2023). This rigorous evaluation process underscores the code's commitment to robust model assessment and refinement.

Continuing the analysis, the code transitions to training a Logistic Regression classifier, leveraging its interpretability and efficiency for binary classification tasks (Dang et al., 2024). Following training, the logistic regression model undergoes evaluation, mirroring the comprehensive assessment process employed for the Random Forest classifier. Classification reports and accuracy scores are computed, providing a comparative analysis of model performance, and aiding in the selection of the most effective classification approach. Confusion matrices for the logistic regression model are also constructed and visualized, enhancing the understanding of classification outcomes, and facilitating targeted model optimization strategies (Singh Kushwah et al., 2021).

Concluding the analysis, Principal Component Analysis (PCA) scatter plots are generated to visualize the distribution of data points in a reduced-dimensional space for "Walking" and "Jogging" activities. These visualizations offer valuable insights into underlying data patterns and relationships, complementing the rigorous evaluation of classification models and providing a holistic understanding of the dataset's characteristics (Dong et al., 2022). In summary, the code's systematic approach to data preprocessing, model training, evaluation, and visualization underscores its effectiveness in analyzing and optimizing classification model performance on the given dataset, contributing to informed decision-making and actionable insights in classification tasks.

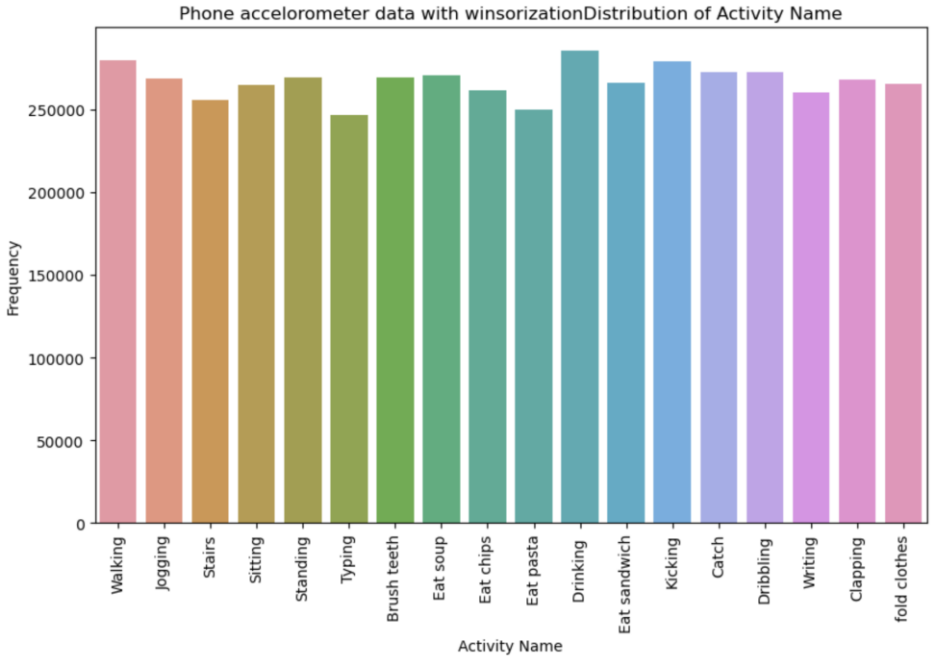
4. Interpretation of Data Visualization

In this section, we interpret the data visualizations, each visualization provides valuable information about the distribution, summary statistics, algorithm training and result, correlation, and behavior of sensor data, aiding in understanding human activities and behavior.

Distribution of Activity Labels

Phone accelerometer data (winsorized) was plotted to understand the distribution of each activity. This visualization allows us to assess the balance or imbalance in activity classes. From Figure 1, there are 18 activities with specific frequency. Activities like typing and eating pasta have lower frequencies, while activities like drinking and walking have higher frequencies. A decision was made to compare high and lower frequencies thereby picking walking and jogging for analysis.

Figure. 1



Summary Statistics

The summary statistics of numerical columns (X, Y, Z) provide insights into the central tendency and spread of sensor readings. For example, mean and standard deviation values indicate the typical magnitude and variability of accelerometer and gyroscope measurements in different axes as shown in Figure 2, by comparing these statistics across activities, we can observe variations in movement intensity and directionality.

```
In [209]: all_data[['X', 'Y', 'Z']].describe()
```

Out[209]:

	X	Y	Z
count	1.563043e+07	1.563043e+07	1.563043e+07
mean	1.061043e-01	-1.553642e+00	5.786174e-01
std	4.360732e+00	4.414240e+00	3.609798e+00
min	-9.220520e+00	-1.101991e+01	-8.133701e+00
25%	-1.264388e+00	-4.016933e+00	-6.402741e-01
50%	-1.968384e-03	-1.290283e-01	4.527374e-03
75%	1.110311e+00	1.213215e-01	1.723975e+00
max	1.176017e+01	9.707890e+00	9.343672e+00

Figure 2

Comparison of classification report of Random Forest and Logistic Regression:

Figure 3, Random Forest classification report

```
Classification Report of Walking & Jogging Activity of Phone Accelerometer using Random Forest:
              precision    recall  f1-score   support

     Jogging       0.67      0.65      0.66     53573
     Walking       0.68      0.70      0.69     56073

    accuracy                           0.67    109646
   macro avg       0.67      0.67      0.67    109646
weighted avg       0.67      0.67      0.67    109646


Accuracy: 0.6749448224285428
```

Figure 3

Logistic regression result (estimates the coefficients of the linear equation by maximizing the likelihood of the data using an iterative algorithm) as shown in Figure 4

```
Classification Report of Walking & Jogging Activity of Phone Accelerometer using Logistic Regression Classificatio
n:
              precision    recall  f1-score   support

     Jogging       0.53      0.44      0.48     53573
     Walking       0.54      0.63      0.58     56073

    accuracy                           0.54    109646
   macro avg       0.54      0.54      0.53    109646
weighted avg       0.54      0.54      0.53    109646


Logistic Regression Accuracy: 0.5374933878116849
```

Figure 4

From Figure 3 and Figure 4, it is observed that Random Forest outperforms Logistic regression, this prompted the decision to use Random Forest for the modeling.

Principal component analysis (PCA) was applied to the Random Forest classifier to visualize the distribution and clustering of the two activities in a two-dimensional space using a sample size of 100.

Figure 5

The plot in Figure 5 shows how walking and jogging activities are distributed in a two-dimensional space based on PCA features. The distinct clustering suggests that the classifier can differentiate between these two activities.

Figure 6 provides insights into how walking activity data is distributed in a two-dimensional space using PCA features with a sample size of 100.

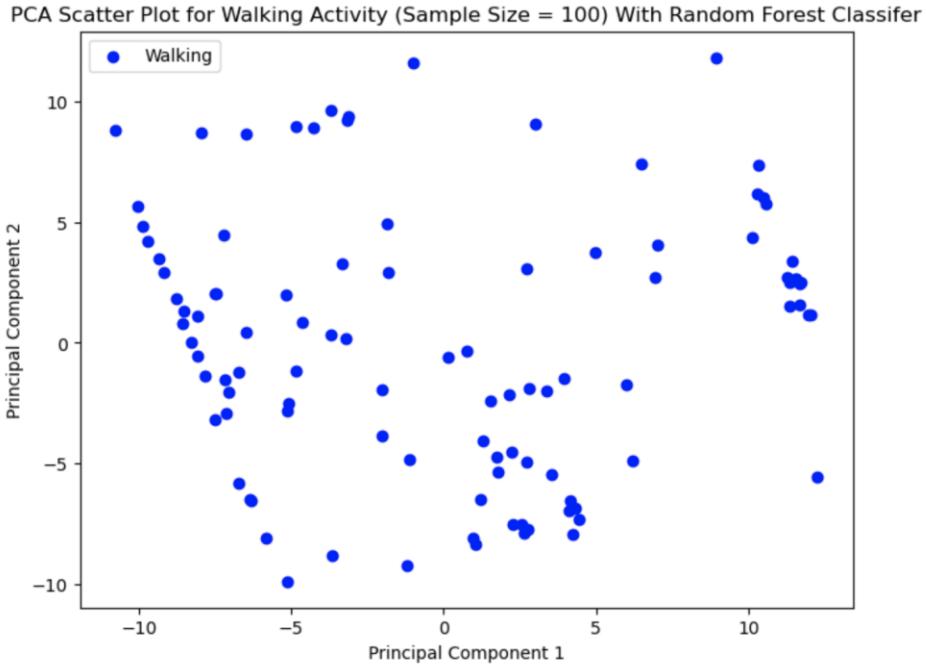


Figure 6

Figure 7 also provides insights into how jogging activity data is distributed in a two-dimensional space using PCA features.

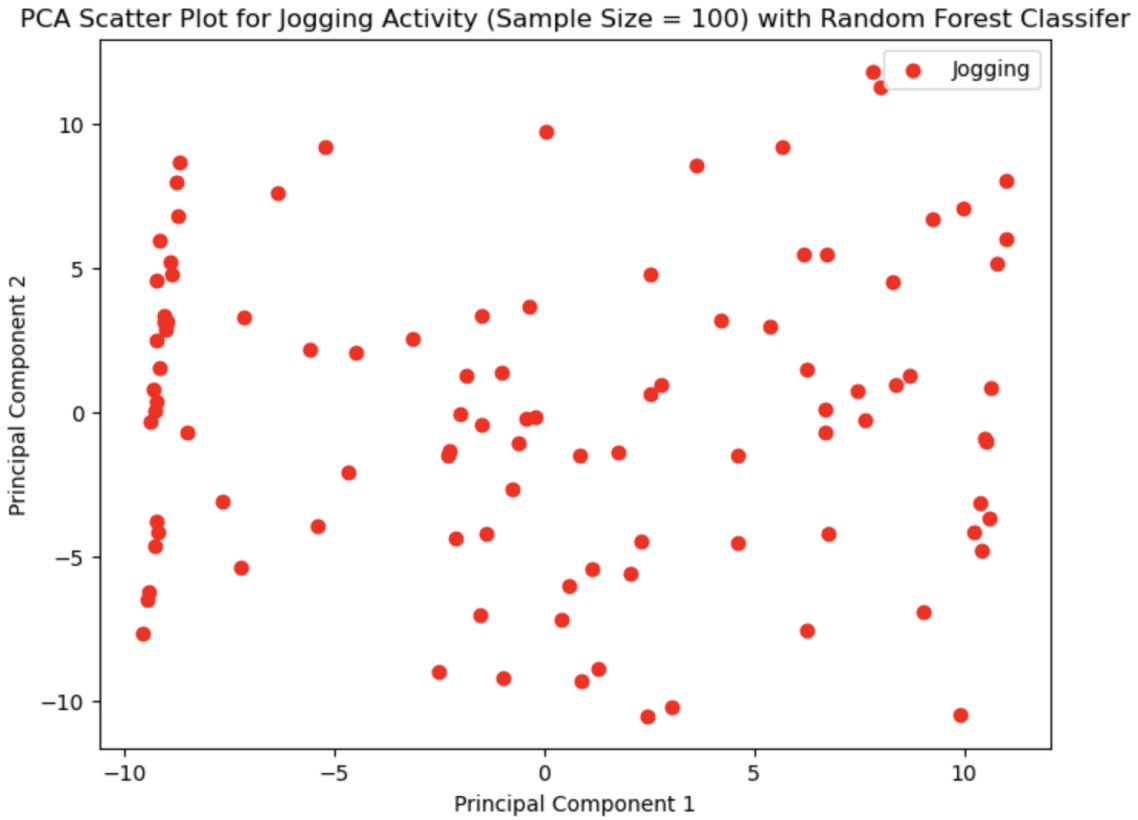


Figure 7

5. Data Mining Solution

In this section, we propose a data mining solution aimed at building a predictive model for accurately classifying human activities based on sensor data from accelerometers and gyroscopes. The approach involves feature selection, model selection, cross-validation, and evaluation metrics tailored to multiclass classification tasks.

The proposed solution includes:

Feature Selection: Extract relevant features from accelerometer and gyroscope data, such as mean, standard deviation, and frequency domain features, to capture underlying activity patterns.
Model Selection: Experiment with various machine learning algorithms like Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM), assessing their performance using metrics like accuracy, precision, recall, and F1-score.
Cross-Validation: Employ k-fold cross-validation to evaluate model generalization performance and mitigate overfitting, tuning hyperparameters using techniques like grid search or random search.
Evaluation Metrics: Assess model performance using appropriate evaluation metrics tailored to multiclass classification tasks, gaining insights into classification accuracy across different activity classes.
The proposed data mining solution incorporates insights from data visualization and understanding of dataset characteristics, emphasizing feature engineering, model training, cross-validation, and model selection for real-time activity recognition applications.

6. Summary

The study of Wireless Independent System Detection and Motion (WISDM) datasets has uncovered crucial insights into motion-based biometrics and data visualization. Employing methodologies like data visualization and machine learning, researchers have found hidden patterns and correlations within the dataset, helping to understand human activities and behavior better.

Implications of Findings:

These findings have important implications for various fields, including healthcare, fitness tracking, and human-computer interaction. Accurately identifying human activities based on sensor data could be useful for applications such as detecting falls in older adults, personalized fitness guidance, and monitoring activities in smart environments. Additionally, the insights from data visualization can help design interfaces and systems that are more user-friendly and intelligent.

Future Directions:

In the future, further research could refine machine learning models and explore more advanced algorithms to improve the accuracy and reliability of activity recognition systems. Long-term studies using real-world data could provide insights into behavior over time, enabling personalized interventions and predictive analysis. Collaborations with experts in psychology, biomechanics, and sensor technology could also deepen our understanding of human emotion and behavior, leading to innovations in assistive technologies and human-centered design.

Overview of Outcomes:

To sum up, this study has shown the effectiveness of data visualization and machine learning in analyzing motion-related datasets. Through careful data analysis and model evaluation, researchers achieved accurate classification of human activities using smartphone and smartwatch sensor data. These models, along with insightful visualizations, offer valuable tools for understanding human behavior and making informed decisions in various fields. Overall, this research contributes to advancing our knowledge in motion-based biometrics and sets the stage for future studies to leverage data-driven insights for improving human well-being and quality of life.

7. Reference

ATHOTA, R.K. and SUMATHI, D., 2022. Human activity recognition based on hybrid learning algorithm for wearable sensor data. *Measurement: Sensors*. 24, p. 100512.

DONG, W., WOZNIAK, M., WU, J., LI, W., and BAI, Z., 2022. De-Noising Aggregation of Graph Neural Networks by Using Principal Component Analysis. *IEEE Transactions on Industrial Informatics*. pp. 1–1.

GARY, W., 2019. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset. UCI Machine Learning Repository. *https://doi.org/10.24432/C5HK59*.

JANNAT, A., MD SHAFIQUL ISLAM, YANG, S.-H., and LIU, H., 2023. Efficient Wi-Fi-Based Human Activity Recognition Using Adaptive Antenna Elimination. *IEEE access*. 11, pp. 105440–105454.

KOSARA, R., 2016. Presentation-Oriented Visualization Techniques. *IEEE Computer Graphics and Applications*. 36 (1), pp. 80–85.

LI, Y., YU, L., LIAO, J., SU, G., HASHMI AMMARAH, LIU, L., and WANG, S., 2022. A single smartwatch-based segmentation approach in human activity recognition. *Pervasive and mobile computing*. 83, pp. 101600–101600.

PEDREGOSA, F., PEDREGOSA@INRIA, F., FR, ORG, G., MICHEL, V., FR, B., GRISEL, O., GRISEL@ENSTA, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., COM, V., VANDERPLAS, J., COM, A., COURNAPEAU, D., VAROQUAUX, G., GRAMFORT, A., THIRION, B., DUBOURG, V., and PASSOS, A., 2011. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay. *Journal of Machine Learning Research* [online]. 12, pp. 2825–2830. Available from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/.

SAKORN MEKRUKSAVANICH, PONNIPA JANTAWONG, NARIT HNOOHOM, and ANUCHIT JITPATTANAKUL, 2022. Heterogeneous Recognition of Human Activity with CNN and RNN-based Networks using Smartphone and Smartwatch Sensors.

SCIKIT-LEARN, 2018. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. *Scikit-learn.org* [online]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

SHAKEEL, H.M., IRAM, S., AL-AQRABI, H., ALSBOUI, T., and HILL, R., 2022. A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research

Developments, Challenges and Future Domain Specific Visualization Framework. *IEEE Access*. 10, pp. 96581–96601.

SHU, X. and YE, Y., 2023. Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*. 110, p. 102817.

SINGH KUSHWAH, J., KUMAR, A., PATEL, S., SONI, R., GAWANDE, A., and GUPTA, S., 2021. Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*.

VALERO-CARRERAS, D., ALCARAZ, J., and LANDETE, M., 2023. Comparing two SVM models through different metrics based on the confusion matrix. *Computers & Operations Research* [online]. 152, p. 106131. Available from: https://www.sciencedirect.com/science/article/pii/S0305054822003616.

WEI ZHONG TEE, DAVE, R., NAEEM SELIYA, and MOUNIKA VANAMALA, 2022. Human Activity Recognition models using Limited Consumer Device Sensors and Machine Learning. *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*.

WEISS, G.M., LOCKHART, J.W., PULICKAL, T.T., MCHUGH, P.T., RONAN, I.H., and TIMKO, J.L., 2016. Actitracker: A Smartphone-Based Activity Recognition System for Improving Health and Well-Being. *IEEE Xplore* [online]. Available from: https://ieeexplore.ieee.org/abstract/document/7796955?casa_token=Td6DjmhcWJEAAAAA:15293M5OtkAkPbyG6qEB3PFKJH-2Y2i3cGmqwaATJqrym5n8QOVYXwRv1T14sLilAsc5n7U6 [Accessed 8 Apr 2023].

WEISS, G.M., YONEDA, K., and HAYAJNEH, T., 2019. Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living. *IEEE Access*. 7, pp. 133190–133202.

ZHENG, J., LIU, Y., and GE, Z., 2022. Dynamic ensemble selection based improved random forests for fault classification in industrial processes. *IFAC Journal of Systems and Control*. 20, p. 100189.

ZHOU, F., WANG, R., SU, H., and XU, S., 2022. A Human Activity Recognition Model Based on Wearable Sensor. Presented at the 2022 9th International Conference on Digital Home (ICDH). pp. 169–174.

8. Appendices

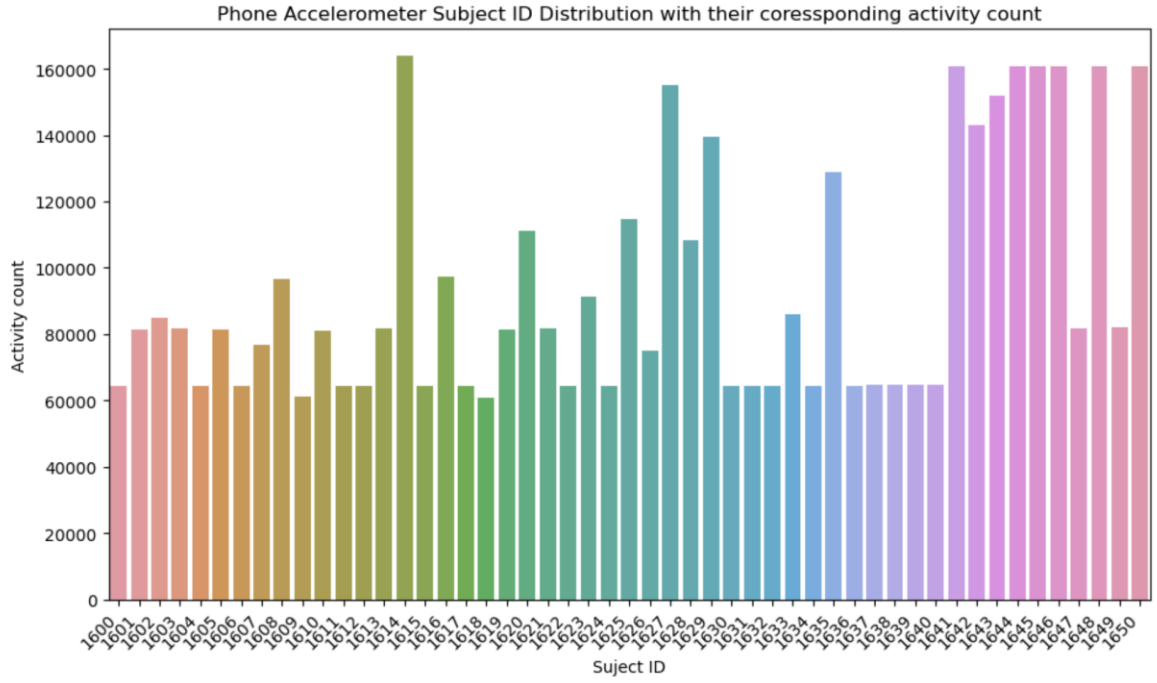
Appendix A: The phone accelerometer data (preview 6000 records).

```
phone_accel_data.head(6000)
```

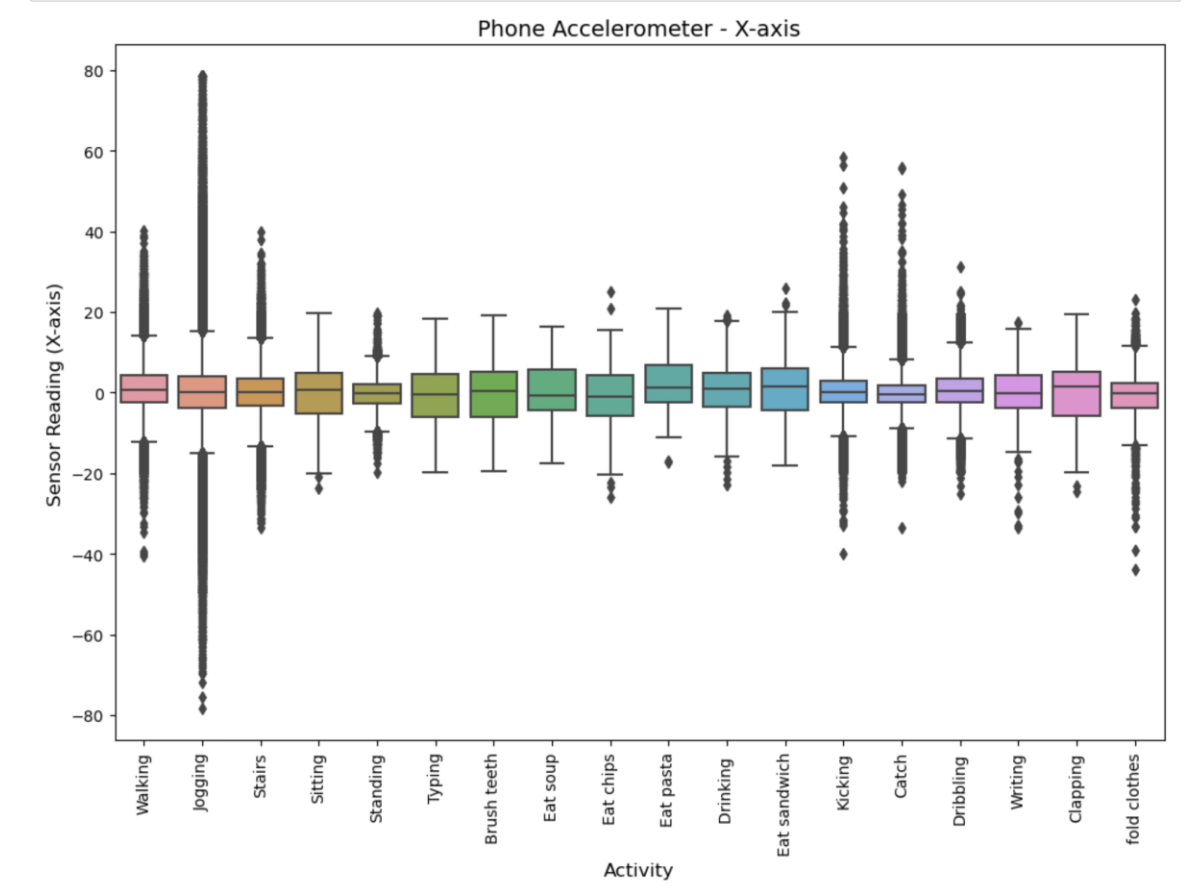
	Subject_ID	Activity_Label	Timestamp	X	Y	Z	Activity_Name	Group_Category
0	1631	A	1970-01-18 23:37:52.620859145	-3.231689	0.960129	1.223938	Walking	Non-Hand Oriented
1	1631	A	1970-01-18 23:37:52.671213149	-3.688065	2.486359	2.971283	Walking	Non-Hand Oriented
2	1631	A	1970-01-18 23:37:52.721567153	-2.923523	8.615723	5.365753	Walking	Non-Hand Oriented
3	1631	A	1970-01-18 23:37:52.771921157	0.362640	16.023514	7.035049	Walking	Non-Hand Oriented
4	1631	A	1970-01-18 23:37:52.822275160	-5.205841	7.684662	6.512863	Walking	Non-Hand Oriented
...	...	...	...	...	...	...	...	...
5995	1631	B	1970-01-18 23:43:20.777897728	-5.850388	-3.668396	-6.985275	Jogging	Non-Hand Oriented
5996	1631	B	1970-01-18 23:43:20.828251732	-4.480866	-4.668900	-2.815567	Jogging	Non-Hand Oriented
5997	1631	B	1970-01-18 23:43:20.878605736	0.759018	9.867371	10.647934	Jogging	Non-Hand Oriented
5998	1631	B	1970-01-18 23:43:20.928959740	14.411881	10.682022	-11.351410	Jogging	Non-Hand Oriented
5999	1631	B	1970-01-18 23:43:20.979313743	8.055176	-1.791367	-11.602173	Jogging	Non-Hand Oriented

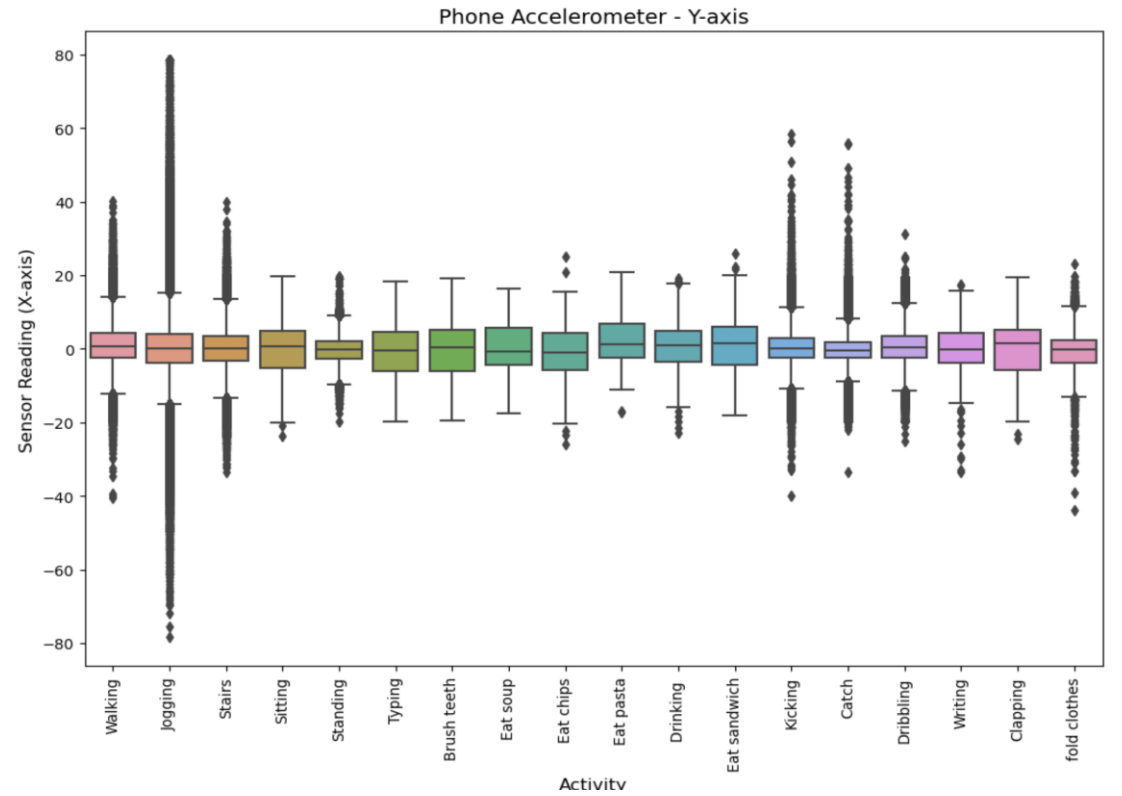
6000 rows × 8 columns

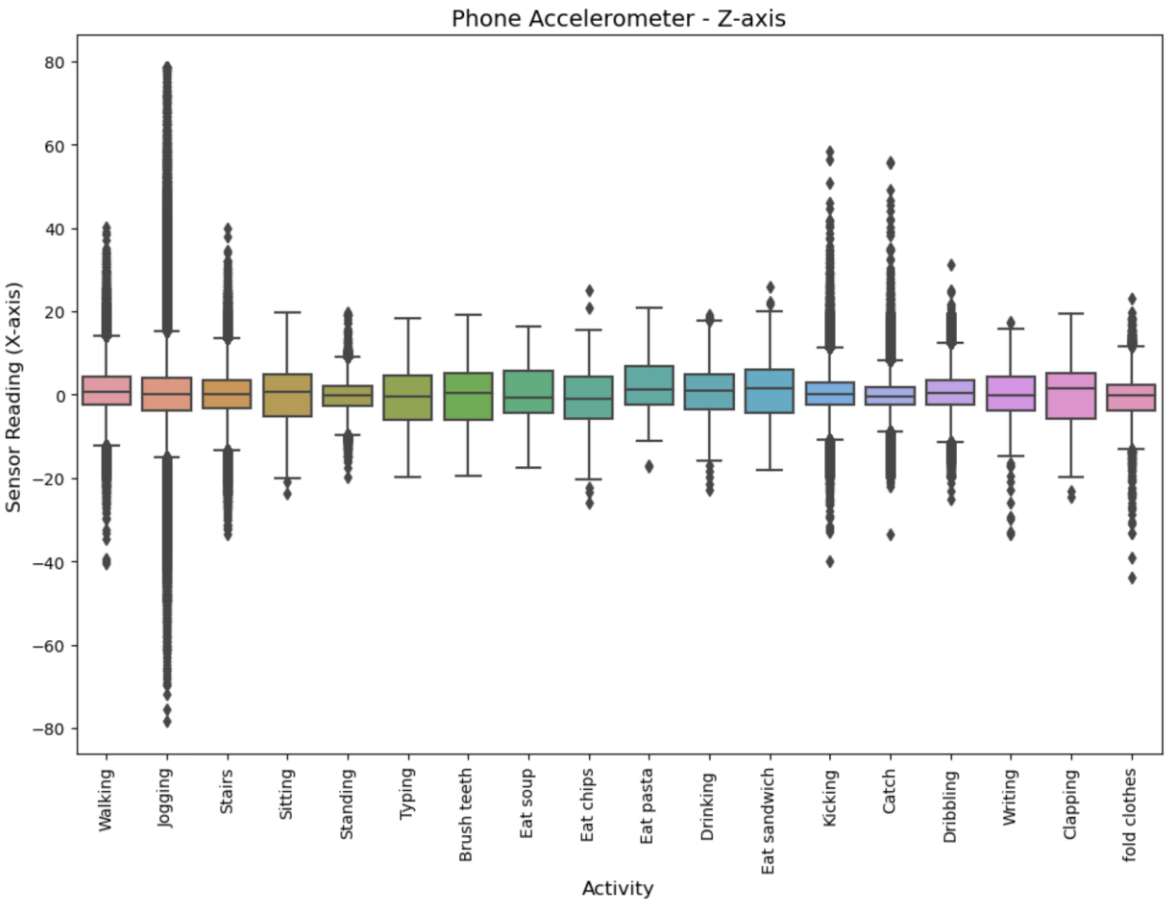
Appendix B: The phone accelerometer data distribution



Appendix C: Phone accelerometer data X, Y, and Z-axis boxplot (to identify outliers) showing all 18 activities.

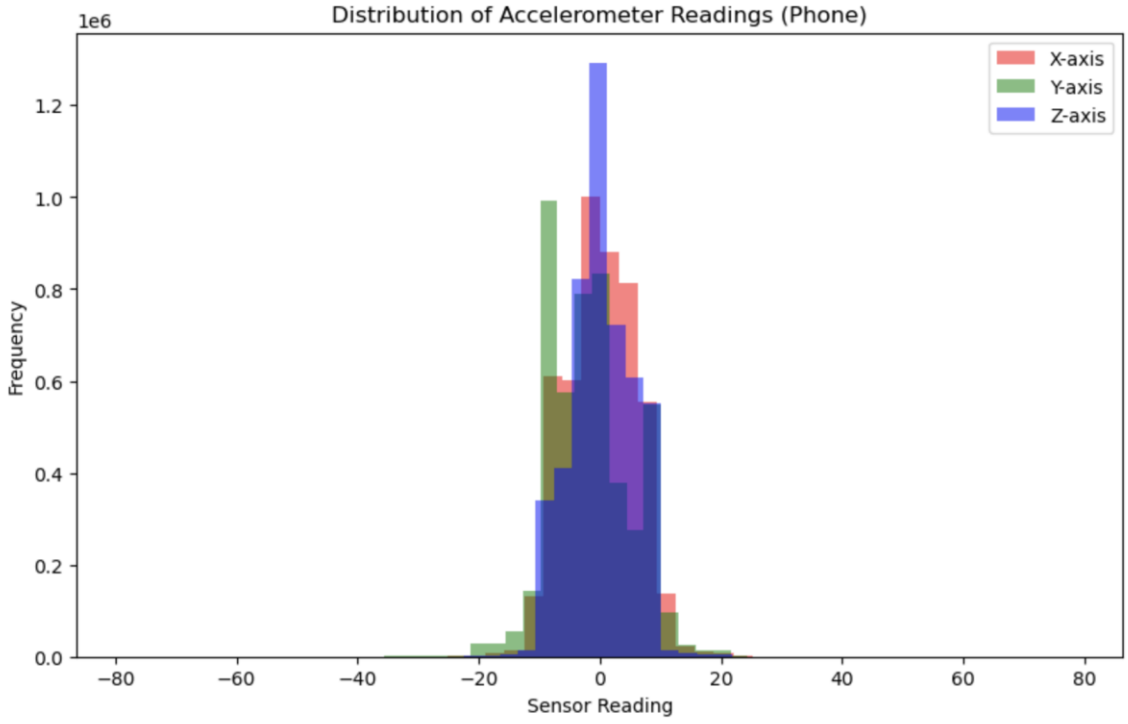




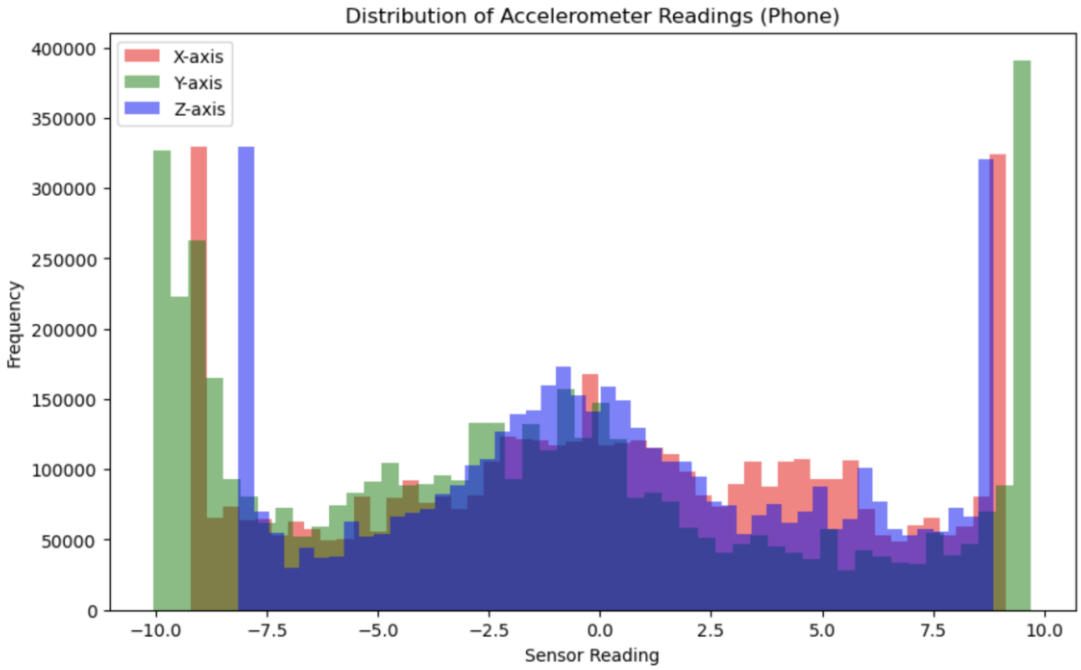


Appendix D: The distribution of phone accelerometer reading before and after applying winsorization.

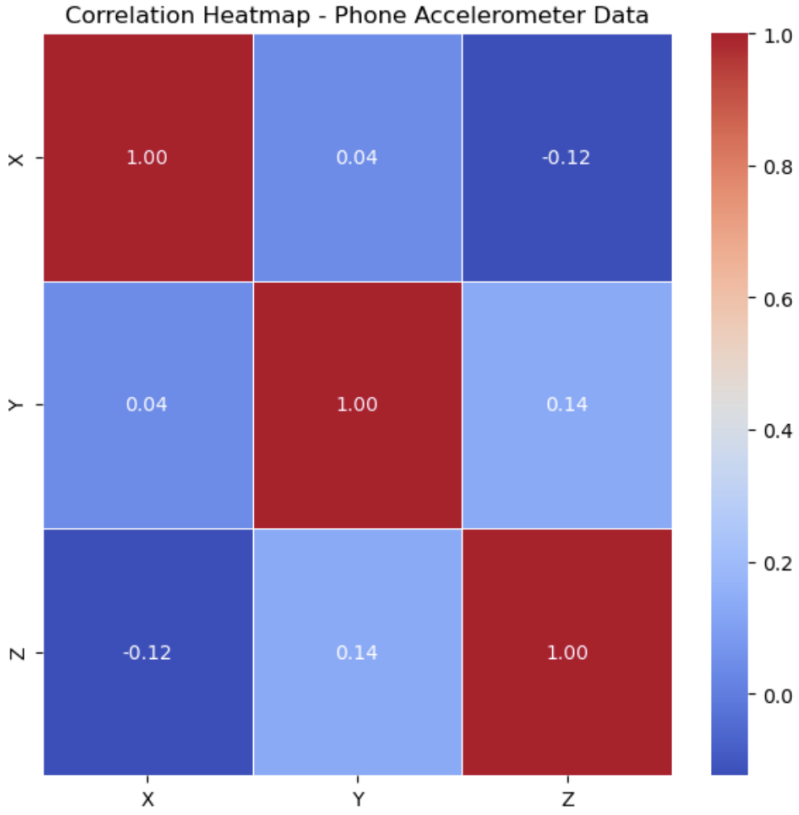
Before Winsorization



After winsorization

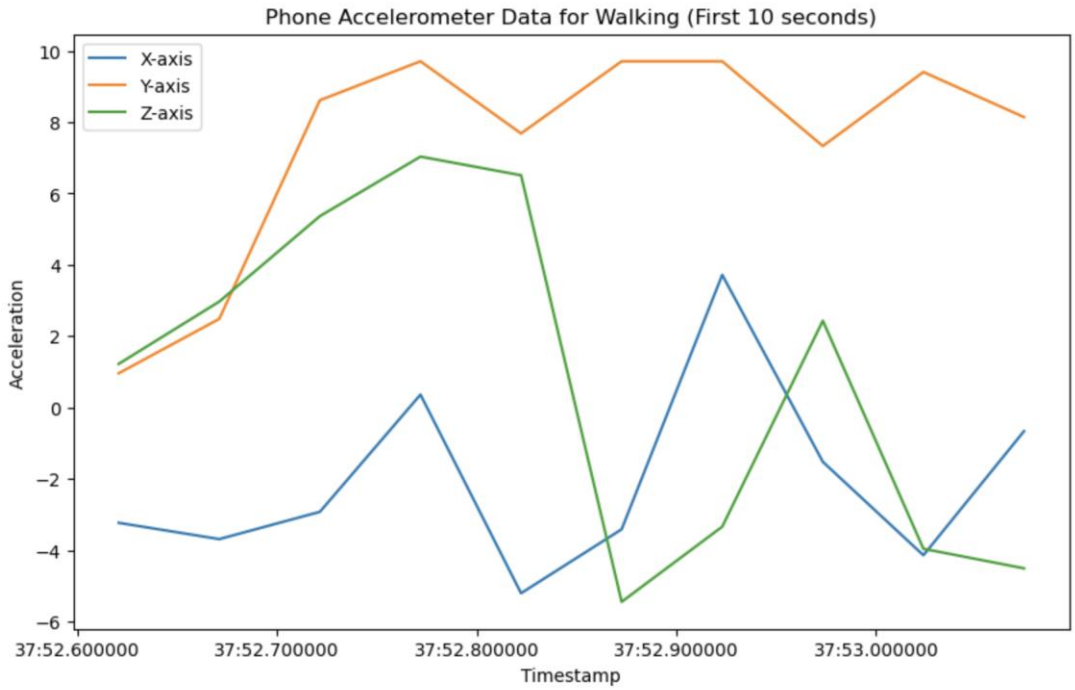


Appendix E: Correlation of phone accelerometer data

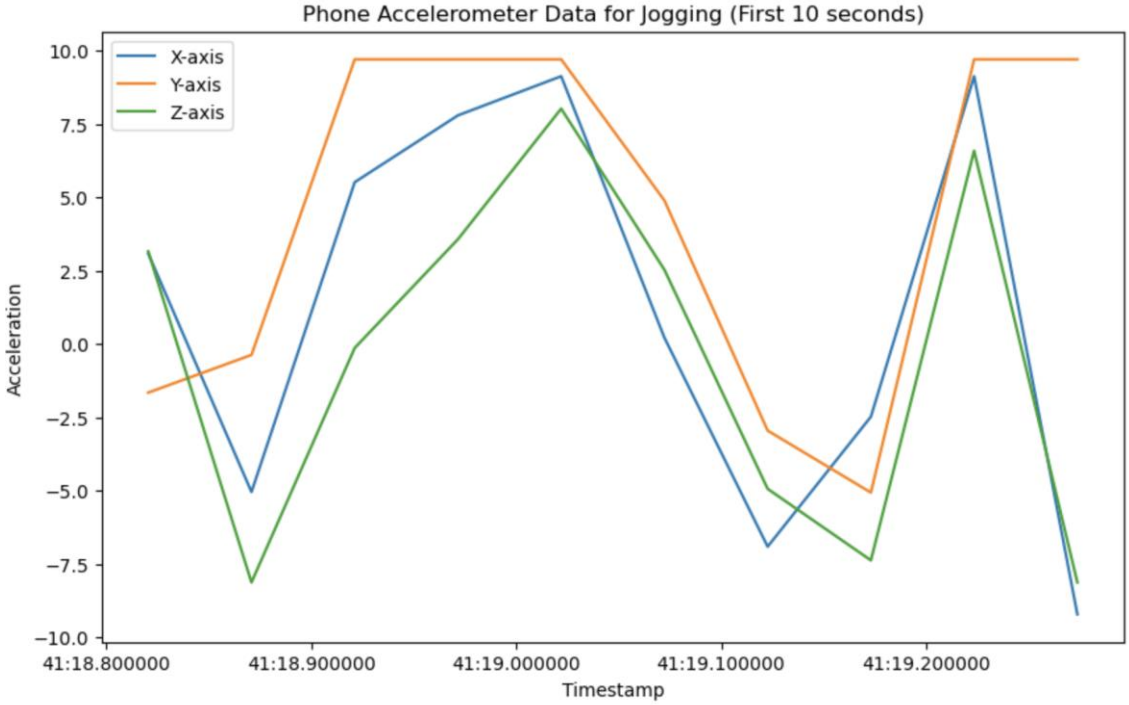


Appendix F: Time series plot for phone accelerometer data (activity = walking and Jogging) for the first 10 seconds.

For Walking



For Jogging



Appendix G: Confusion matrix provides insights into how well the Random Forest model with PCA performed in classifying jogging and walking activities.