Analyzing Behavioral Patterns Over Time for User Authentication Using Apple Watch Accelerometer Data

Maharage Nisansala Sevwandi Perera[1,a)] Allam Shehata[2,b)] Chi Xu[2,c)] Xiang Li[2,d)] Franziska Zimmer[1,e)] Ryosuke Kobayashi[1,f)] Mhd Irvan[1,g)] Rie Shigetomi Yamaguchi[1,h)] Yasushi Yagi[2,i)]

Abstract: In recent years, behavioral authentication has garnered attention as a promising candidate for non-intrusive authentication mechanisms. In behavioral authentication, the user's current behavioral patterns are compared to their historical behavioral profile. However, a key challenge in behavioral authentication is the impact of behavioral changes over time, which may affect the accuracy of correctly identifying legitimate users. Such changes can result from environmental factors, emotional states, fatigue, or the passage of time. This paper analyzes the behaviors of 100 Apple Watch users based on accelerometer data collected over three hours of experimentation. We trained the model using two different classifiers: Random Forest and XGBoost, and evaluated their authentication performance using the last 40–60 minutes of data. Authentication models with both classifiers accepted genuine users with high confidence. Despite behavioral variations, such as those induced by fatigue, our findings indicate that behavioral patterns remain a viable and practical basis for continuous authentication.

Keywords: behavioral authentication, cross-time validation, behavioral changes, Apple Watch, accelerometer data

1. Introduction

The concept of behavioral authentication has evolved over time, dating back to the 1980s or earlier. Behavioral biometrics such as keystroke dynamics (typing patterns) were among the earliest forms of behavioral authentication [1]. The term "behavioral authentication" and its application in the context of wearable and mobile sensors gained popularity in the 2010s, driven by the rapid development of wearables and significant contributions from academics and industrial researchers [2], [3]. In behavioral authentication, a user profile built based on the behavioral patterns of the user is employed to validate the user at the time of authentication. Thus, the current behavioral pattern is cross-checked against past behavioral patterns without requiring the user to input any credentials, making behavioral authentication a non-intrusive authentication method. Due to its passive nature, behavioral authentication is widely regarded as a promising approach for continuous authentication, helping to mitigate risks such as impersonation and identity theft [4].

Behavioral biometrics can include keystroke patterns, gait, accelerometer data, GPS movement patterns, or other user-specific behaviors [5]. For example, in virtual reality and online gaming platforms, not only physical movements such as hand gestures but also decision-making patterns in gameplay can serve as behavioral cues [6]. On touchscreen devices, features such as swipe direction, dragging motion, and touch pressure and speed are commonly analyzed as behavioral signals [7]. Among all behavioral data sources, wearable devices are among the most promising. Devices like smartphones and smartwatches are widely adopted and capable of capturing real-time behavioral signals such as step count, movement patterns, and distance walked [8], [9]. For instance, while Chen et al. [10] showed that step count data can be used to authenticate users, Buriro et al. [11] proposed a bi-modal authentication method for smartwatch authentication.

However, behavioral authentication still faces significant challenges, as user behaviors can vary under different circumstances, making it more difficult to verify genuine

[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2] D3 Center, The University of Osaka, Japan
[a)] perera.nisansala@yamagula.ic.i.u-tokyo.ac.jp
[b)] allam@yy.d3c.osaka-u.ac.jp
[c)] xu@yy.d3c.osaka-u.ac.jp
[d)] li@yy.d3c.osaka-u.ac.jp
[e)] zimmer@yamagula.ic.i.u-tokyo.ac.jp
[f)] kobayashi@yamagula.ic.i.u-tokyo.ac.jp
[g)] irvan@yamagula.ic.i.u-tokyo.ac.jp
[h)] yamaguchi.rie@i.u-tokyo.ac.jp
[i)] yagi@yy.d3c.osaka-u.ac.jp

users reliably. For example, Okawa et al. [12] showed that individuals often exhibit multiple lifestyles depending on the day of the week, and proposed a two-template authentication approach to address this variability. Similarly, a profile created when the user is in good health may fail to match their behavior when they are fatigued or unwell, as their activity patterns may slow or become less dynamic [13].

Therefore, it is crucial to analyze behavioral changes over time and explore strategies to mitigate authentication errors that arise from such variations in user behavior.

Contribution

This study analyzes the behavioral changes of 100 Apple Watch users over time and demonstrates that, with careful feature selection and appropriate model training, users can still be distinguished and authenticated with high accuracy despite behavioral variations. Our results provide encouraging evidence for the feasibility of behavioral authentication and highlight potential directions for future research in enhancing its robustness.

2. Our Approach

This section gives an overview of our method along with the data collection process. Moreover, it explains the feature extraction process and the classifiers selected.

2.1 Overview of the Method of this Study

This paper analyzes behavioral changes over time in a cohort of 100 Apple Watch users. Specifically, we focus on accelerometer data collected using the SensorLog app [14] installed on iPhones paired with the Apple Watches. Each participant took part in a pre-configured experimental session lasting approximately three hours. Since three hours is relatively long, we assume that participants experienced fatigue during the final 40–60 minutes. This period is of particular interest, as we investigate whether users can still be authenticated under fatigue. To this end, we employed two classifiers, namely, Random Forest and XGBoost, and trained two separate models using accelerometer data from the first two hours of the session. We then evaluated these models on data from the fatigue period. Our analysis revealed that, despite observable behavioral changes over time, user authentication can still be performed with high accuracy. These findings confirm that behavioral patterns remain a viable and practical feature for continuous authentication, even under conditions of physical fatigue.

2.2 Data Collection

To analyze the behavioral patterns of Apple Watch users for authentication based on their behavior, we utilize a dataset collected while participants moved within a confined experimental space and participated in tasks such as iris scanning, face recognition, and the word-chain game (shiritori). These experimental modalities did not follow a fixed sequence, and their data were not used in this study.

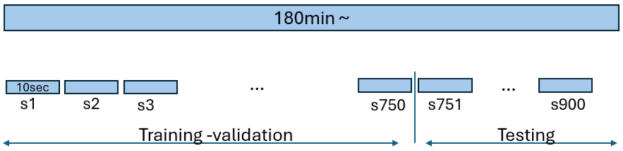


Fig. 1: Slicing user data into 10-second slices/intervals

Instead, we focus solely on Apple Watch sensor log data recorded while participants were in the experimental area.

Each participant was provided with an Apple Watch (Series 8), which was connected to an iPhone SE and worn from the beginning of the session until they left the experimental space. The sensor log recorded data of 60 features, and the full dataset includes over 100 participants (82 females and 28 males). For this first step toward applying behavioral biometrics for user authentication, we consider only the accelerometer data from 100 participants. The accelerometer readings consist of timestamps and acceleration values along the X, Y, and Z axes. No personally identifiable information was collected or recorded.

Data collection lasted approximately three hours per participant. The SensorLog application [14] was installed on both the iPhone and Apple Watch to record and stream raw sensor and health data. SensorLog was configured with a logging rate of 20 Hz for all enabled sensors, allowing for potential downsampling. All available motion sensors, including the accelerometer, gyroscope, and magnetometer, were activated.

2.3 Preprocessing

Each user's data spans approximately three hours, equivalent to 180 minutes or 10,800 seconds. Since this duration is relatively long, we divided the data into 10-second time slices. Depending on the user's exact participation time, this results in approximately 900–1,080 slices per user. As illustrated in **Fig. 1**, we used the first 750 slices for model training and validation, reserving the remaining later slices for testing, i.e., for authentication using previously unseen data.

For model training, validation, and authentication, we conducted our experiments for increasing user groups from 10 to 100, with an increment of 10. For the training and validation step, we first generated balanced batch pairs with up to 100,000 pairs per class. To construct these, we collected all consecutive slice pairs from the same user in the concern group. Then, to balance the dataset, we generated an equal number of negative pairs by combining slices from two different users who show deceptively similar behavior. We use cosine similarity to check the similarity score of two pairs to find negative pairs that are hard to distinguish (hard-negative pairs). Since the number of possible pairs can become very large, we limited the maximum number of pairs per class to 100,000.

2.4 Feature Extraction

Even though all available motion sensors, including the accelerometer, gyroscope, and magnetometer, were activated in the Apple Watches worn by the users, this study utilizes accelerometer data only, as it was recorded without any missing values and can be easily used to distinguish user behaviors. Accelerometer sensors are 3-dimensional, i.e., X, Y, and Z. We extracted five statistical features, namely Mean, Standard Deviation, Minimum, Maximum, and Median, from each axis, along with the range and energy of each axis. Moreover, signal magnitude area (SMA) is used. Pairwise correlations between axes (XY, XZ, YZ) were also included to capture the relationship between axes, as they reflect the coordination of the movement that can be unique to a person. Additionally, the magnitude of the axes was calculated, and the above five statistical features, along with the range and energy of magnitude, were considered. Thus, 32 features were considered. However, to capture the changes over time, delta features, i.e., the difference between corresponding features from two consecutive batches, were calculated and used in our research.

The importance of the features to identify the users can vary based on the user and the environment. Feature importance is discussed in Section 3.1.

2.5 Classifiers

To understand behavioral patterns over time, this work employs Random Forest (RF) and XGBoost (XG) as classifiers. RF is a traditional classifier, and the final result of RF is the most-voted class from the forest (many trees) [15]. XGBoost is also a voting classifier, but it uses the gradient-boosting framework to deliver results [16]. We utilize both classifiers and deliver two authentication models based on them. Thus, we compare the performance of classifiers for authenticating users in the created models with increasing user counts (10, 20,..., 100).

3. Distinguishing and Authenticating Users

This section explains the process of distinguishing users and the authentication process. First, important features that affect distinguishing users are discussed. Then, model training-validation and the process of authenticating genuine users are discussed.

3.1 Feature Importance

Important features were examined for both classifiers as the number of users increased. **Fig. 2** with feature importances shows that the ranking of features selected by RF differs somewhat from those selected by XG for the group of 10 users. This indicates that the set of important features may vary depending on the model. Furthermore, the importance of features may also change as more users are added, since the behavior of newly introduced users influences the feature space. The differences in rankings shown in **Fig. 3**, compared to Fig. 2, further confirm this observation.

However, the results show that extracted features such as 'delta_energy_mag', 'delta_median_X', 'delta_median_Z', 'delta_mean_X', 'delta_median_mag', 'delta_range_Z', 'delta_range_X', 'delta_range_Y', and 'delta_mean_mag' consistently appear among the top-ranked features in both models with classifiers RF and XG. These features, which primarily represent energy, ranges, and medians, are commonly identified as important across the models. In addition, rather than individual aggregated features, the differences in feature values between slice pairs, i.e., delta features, are significant, as the model checks for behavioral similarities based on the similarities (or differences) between the two slices given.

3.2 Training and Validation

Models are trained using positive and negative pairs as inputs, enabling the model to identify users during the validation process. Since a balanced dataset with an equal amount of 'same user' (positive data) and 'different users but very similar' (hard-negative data) was generated during the preprocessing stage, the final dataset is suitable for training robust verification models. We trained the model with 10 different groups, each comprising 10, 20, ..., 100 users. We ensure that the set of previous users is included in the next user group. For example, the first 10 users are included in the next 20-user group, those 20 users are included in the next 30-user group, and so on.

First, we selected four users to investigate their behavior in detail throughout the training-validation phase and authentication with unseen data (testing). Those users are 'user_1', 'user_20', 'user_41', and 'user_87', and we selected these users as they showed different accelerations as shown in **Fig. 4**. For instance, while 'user_1' showed very low mean acceleration like -0.32 in X direction (negative acceleration), 'user_87' showed high mean acceleration like 0.38 in X direction. On the other hand, while 'user_20' showed intense acceleration, 'user_41' showed opposite behavior to 'user_20'.

For the training and validation of models, all the generated pairs were first arranged in chronological order to preserve the temporal sequence, and the first part of the data sequence was used for training and later for validation. In other words, data splitting for training and validation is handled by the TimeSeriesSplit cross validator [17]. We selected TimeSeriesSplit as it is more suitable for sequential data, such as our dataset, which contains accelerometer signals. On the other hand, since the training set always consists of earlier observations, while the validation set consists of later observations, the model is evaluated in a way that mimics the real-world deployment. We use a five-fold split such that Fold-1 trains on the first portion of the data and validates on the immediately following portion. At the same time, Fold-2 extends the training window further into time and gains validation on the next batch, so on as shown in **Fig. 5**.

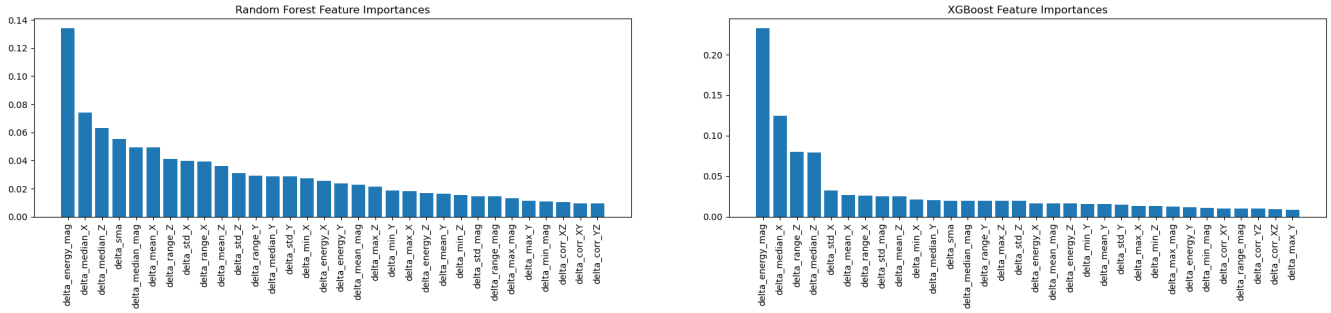


Fig. 2: Feature importance comparison for 10-users group using Random Forest (left) and XGBoost (right).

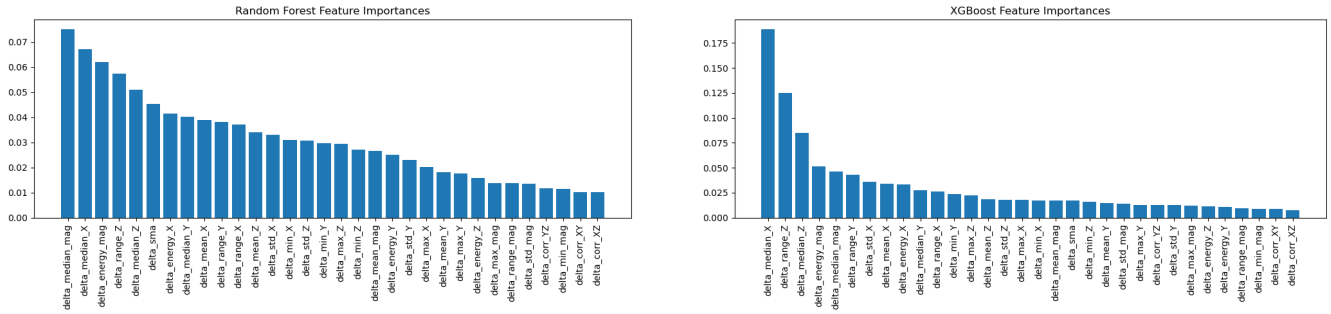


Fig. 3: Feature importance comparison for 20-users group using Random Forest (left) and XGBoost (right).

This strategy ensures that no data leakage occurs from the future into the past, thereby satisfying a critical point in authentication tasks based on sensor data.

With the increasing number of users from 10 to 100, both RF and XG classifiers are employed to train the model separately and evaluate the validation performance incrementally by 10. **Fig. 6** shows the changes in accuracy and other performance matrices with the increasing number of users for the RF and XG classifiers, respectively. These results are the average of distinguishing each user in the considered group. For instance, at the point of 10 (number of users), accuracy and performance matrices are of the average value of distinguishing each user (mean of 10 users).

Both models, RF and XG, exhibit nearly the same trend for overall accuracy and performance metrics as the number of users in the groups increases, except in the last couple of groups (90 and 100), where minor deviations are observed. Interestingly, at the point of 30 users, both models show a sudden drop in Recall, F1-score, and overall Accuracy. This behavior is likely due to the introduction of new users with similar behavioral patterns, making it more difficult for the models to distinguish them. In terms of Precision, the two models differ slightly: the RF model shows a small drop at 60 users, while the XG model shows a slight drop at 50 users. Despite these fluctuations, both models consistently maintain high levels of performance, with overall accuracy and metrics staying above 85% across all user group sizes.

3.3 Authenticating Genuine Users

For the process of authenticating genuine users, unseen data from the remaining slices, specifically from slice number 751, is employed for each user. For each case, two consecutive slices with high slice numbers (greater than 750) are selected, and the models are tested to determine whether they can correctly identify that both slices belong to the same user, thereby confirming genuine user authentication.

In this work, the probability score output by the model, which provides a continuous confidence measure ranging from 0 to 1, is employed. A higher probability indicates more substantial confidence that the authentication attempt is genuine, while a lower probability suggests that it is more likely an imposter attempt. Thus, in this study, we relied on the default probability threshold of 0.5 to make the decision. That is, if the probability score (confidence) is equal to or greater than 0.5, the slices are classified as belonging to a genuine user; otherwise, the attempt is considered an adversarial (imposter) attack.

Based on the above setup, for user authentication, we received the results shown in **Fig. 7**, indicating the overall acceptance rate as the number of users increases. These figures specifically show the overall percentage of models that provide access to genuine users when they are trained with different groups. Each user's acceptance rate is calculated when trained with the respective user group, and Fig. 7 depicts the averaged acceptance rate against increasing user group size.

Both models demonstrate a high probability (greater than 85%) of accepting genuine users, indicating that they perform well in authenticating genuine users even as the group size increases. However, while the XG classifier shows more consistent growth in genuine user authenti-

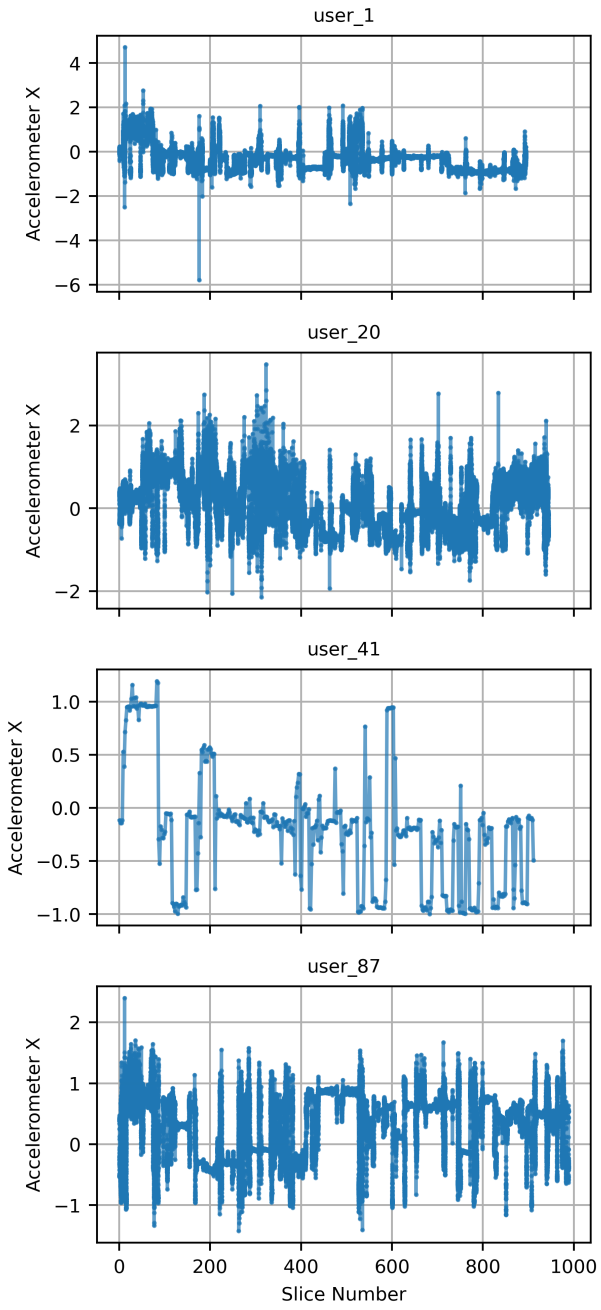


Fig. 4: Accelerometer behavior of selected users

cation compared to the RF classifier, the RF classifier appears to struggle more in reliably judging genuine users. Results indicate that XG is a better candidate than RF with the increasing population to apply in behavioral authentication for our study area.

4. Discussion

This section discusses the performance of models when genuine user access and adversarial access are present. The limitations of this study and the need for further investigation are then discussed.

4.1 Authentication Accuracy Over Time

In this subsection, we further investigate four users

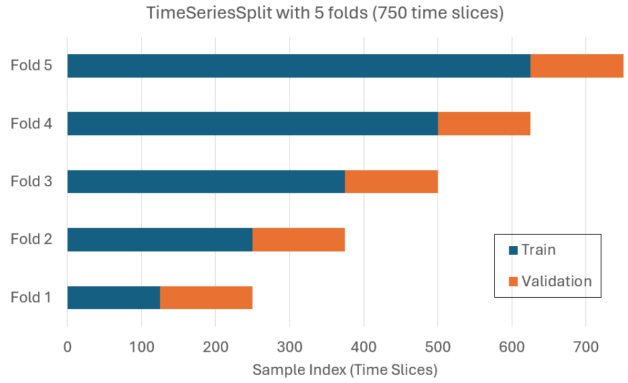


Fig. 5: TimeSeriesSplit with 5-folds for training and validation

('user_1', 'user_20', 'user_41', and 'user_87') selected based on their accelerometer variations.

Figures **??**–**??** illustrate the authentication performance of each user over time when the RF model is applied to the 10-user group, while the same trend is shown in Figures **??**–**??** for the XG model.

From the results, we observe that both models exhibit performance degradation for almost all users near the end of the task. For example, the RF model failed to authenticate 'user_1' around slices 820 and 870 (with slice 870 occurring just before the end of the task). Similarly, it failed to authenticate 'user_87' after slice 950 but before the task ended. The XG model also showed a somewhat similar trend.

These observations suggest that users may have experienced fatigue during the long task. Toward the end, their behavioral patterns, such as reduced acceleration, became more similar, making it more difficult for the models to distinguish them. However, this assumption requires further investigation, which we plan to address in future research.

4.2 Adversarial Access

While maintaining a confidence threshold of 0.5, we conducted a preliminary investigation to evaluate the models' performance against adversarial attacks. The results, however, revealed that it is necessary to adjust the threshold to a stricter level, such as 0.75 or higher, since the models tended to accept impostors as genuine users under the default setting. For this preliminary study, slice pairs were generated by combining one slice from an authentic user with another slice from a different user to simulate imposter access attempts.

In addition to modifying the confidence threshold, another possible approach is to train the models using smaller time slices. This would enable the models to capture finer-grained behavioral details, potentially enhancing their ability to distinguish between genuine and imposter attempts.

4.3 Limitations and Future Work

In addition to the limitations discussed above, several

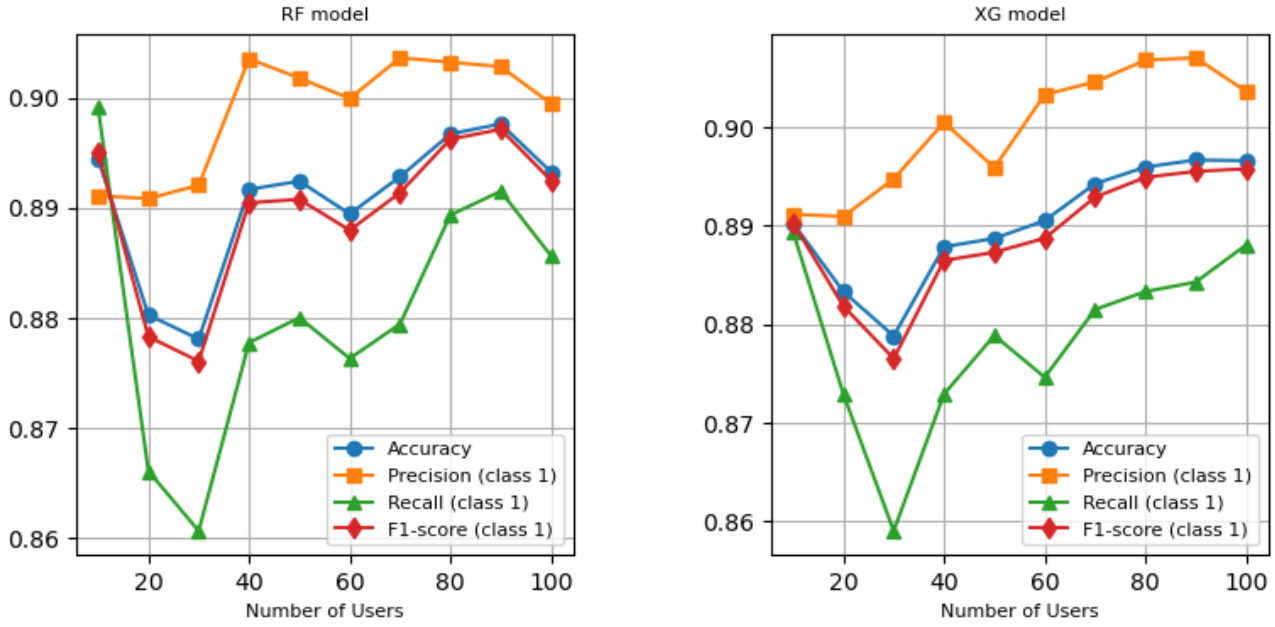


Fig. 6: Overall Accuracy and Performance changs with increasing number of users for Random Forest (left) and XGBoost (right).

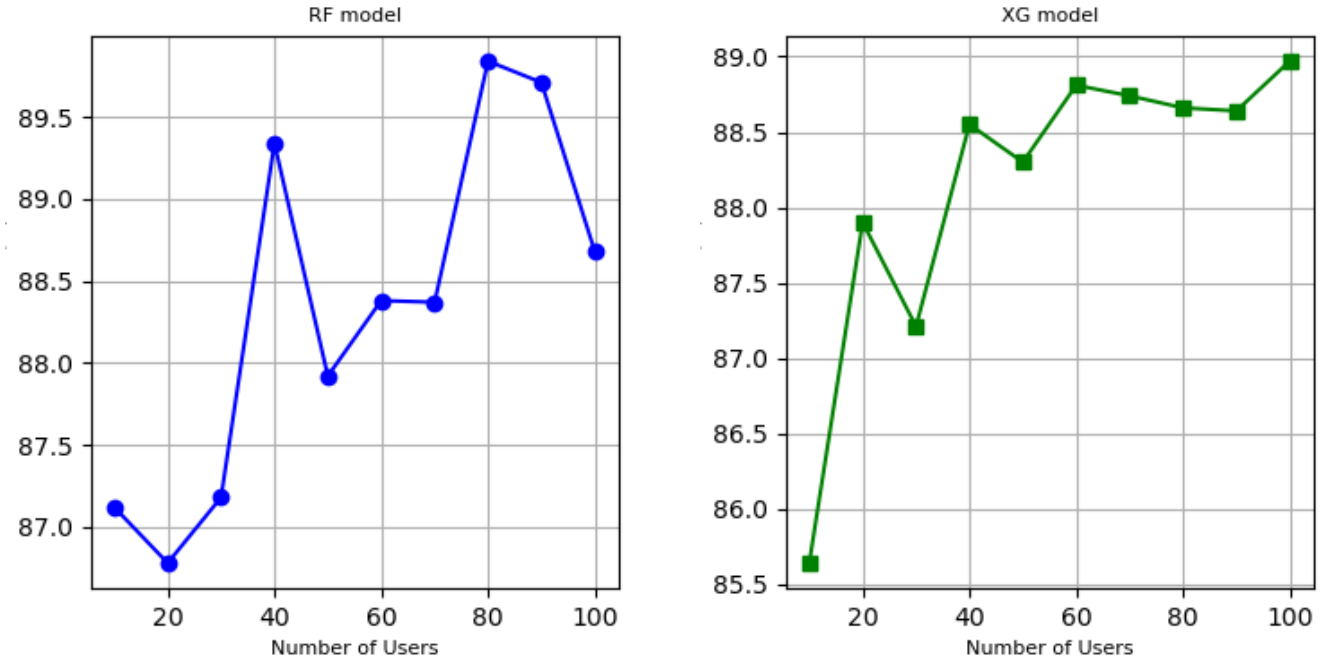
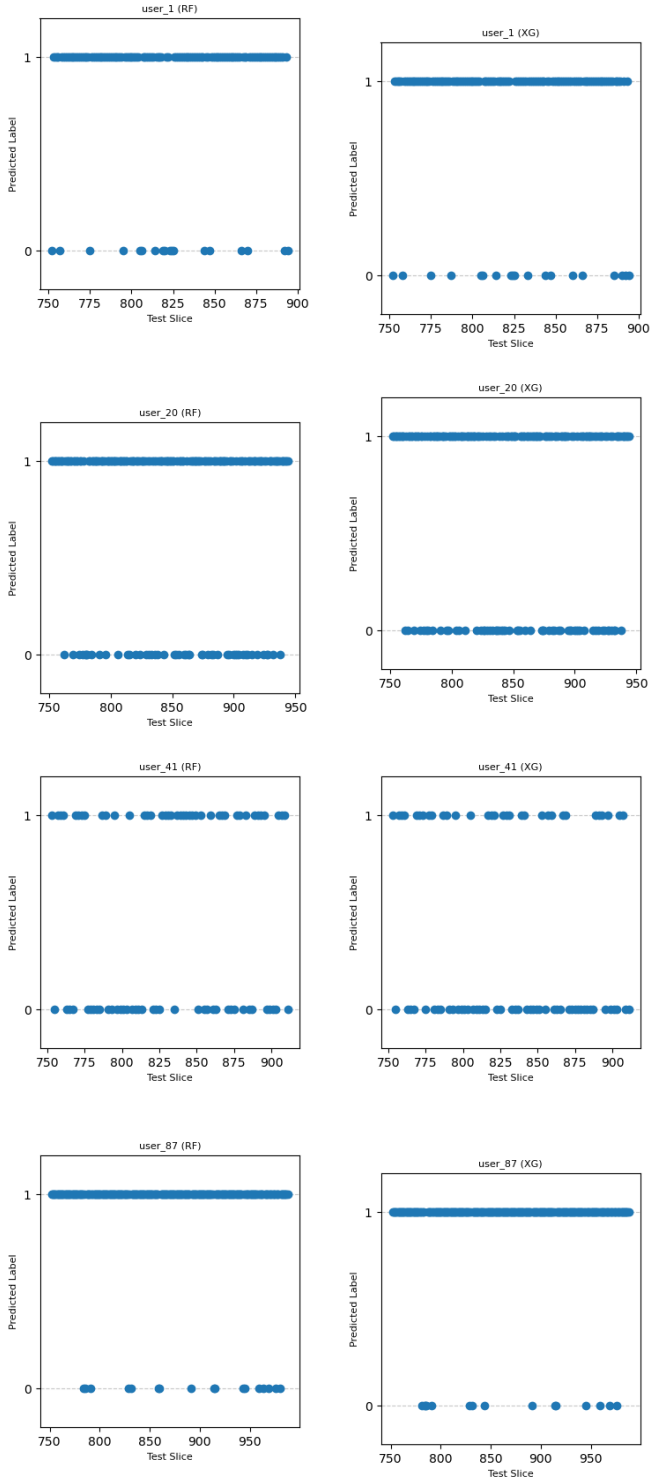


Fig. 7: Overall Acceptance rate of Genuine users Random Forest (left) and XGBoost (right).

further constraints of this study should be noted.

First, this research employed only accelerometer data, meaning that the RF and XG models were effectively mono-modal. However, the Apple Watch provides additional features such as pedometer data, position data, altitude data, and health-related information like walking distance, which could also be investigated to build a more comprehensive authentication model. In this study, we focused on accelerometer data because the other modalities contained missing values that would have required suitable interpolation techniques. In future work, the approach presented here can be expanded by incorporating these additional features to enhance robustness and accuracy.

Second, the RF and XG classifiers rely on aggregated values within each slice, which may limit their ability to construct a highly effective authentication model. Because these models do not directly compare raw slices, they may



(a) Caption 7

Fig. 8: Overall comparison of eight experiments. Each subfigure shows different settings/results.

overlook subtle temporal dynamics in user behavior. To address this limitation, future research should consider sequence-based models such as LSTMs or Transformers, which are well-suited for capturing temporal dependencies in sequential data. Extending the ideas presented in this paper with such models could lead to more robust and accurate authentication systems.

5. Conclusion

This paper analyzed the behavioral patterns of Apple Watch users by employing accelerometer data stored in sensor logs. It utilized two classifiers, RF and XG, to develop an authentication model based on behavioral biometrics. Although the study achieved highly accurate authen-

tication for genuine users, certain limitations remain, such as the need to address adversarial attacks and to explore additional features and models for improved performance. These limitations highlight directions for future extended work. Nevertheless, the current results demonstrate that models trained on paired data slices can effectively distinguish users based on accelerometer data with high overall accuracy, achieving performance levels above 87% for target user classification. These findings indicate a promising foundation for future research on more robust and effective behavioral authentication methods.

Ethics Statement

This study was reviewed and approved by the Ethical Review Board at SANKEN, The University of Osaka, Japan, under protocol number R6-02. All participants provided informed consent prior to their participation, and the study was conducted in accordance with the relevant institutional guidelines.

Acknowledgments This work was supported by JST Moonshot R&D Grant Number JPMJMS2215.

References

[1] Gaines, R. S., Lisowski, W., Press, S. J. and Shapiro, N.: Authentication by keystroke timing: Some preliminary results, Technical report (1980).

[2] Feng, T., Liu, Z., Kwon, K.-A., Shi, W., Carbunar, B., Jiang, Y. and Nguyen, N.: Continuous mobile authentication using touchscreen gestures, *2012 IEEE conference on technologies for homeland security (HST)*, IEEE, pp. 451–456 (2012).

[3] Roth, J., Liu, X. and Metaxas, D.: On continuous user authentication via typing behavior, *IEEE Transactions on Image Processing*, Vol. 23, No. 10, pp. 4611–4624 (2014).

[4] Oduri, S.: Continuous authentication and behavioral biometrics: Enhancing cybersecurity in the digital era, *International Journal of Innovative Research in Science Engineering and Technology*, Vol. 13, No. 7, pp. 13632–13640 (2024).

[5] Oak, R.: A literature survey on authentication using Behavioural biometric techniques, *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*, Vol. 673, Springer, pp. 173–181 (2018).

[6] Zimmer, F., Irvan, M., Perera, M., Tamponi, R., Kobayashi, R. and Shigetomi Yamaguchi, R.: Player Behavior Analysis for Predicting Player Identity Within Pairs in Esports Tournaments: A Case Study of Counter-Strike Using Binary Random Forest Classifier (2025).

[7] Liu, L. and Huang, S.: Dominate and Non-dominate Hand Prediction for Handheld Touchscreen Interaction, *2020 13th International Conference on Human System Interaction (HSI)*, IEEE, pp. 56–62 (2020).

[8] Chen, Z., Shi, K. and Sun, W.: Step Count Print: A Physical Activity-Based Biometric Identifier For User Identification and Authentication, *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2024).

[9] Fuller, D., Anaraki, J. R., Simango, B., Rayner, M., Dorani, F., Bozorgi, A., Luan, H. and Basset, F. A.: Predicting lying, sitting, walking and running using Apple Watch and Fitbit data, *BMJ Open Sport & Exercise Medicine*, Vol. 7, No. 1, p. e001004 (2021).

[10] Chen, Z., Shi, K. and Sun, W.: Walking to authenticate: Identifying robust behavioral biometrics from step count data, *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, IEEE, pp. 393–396 (2023).

[11] Buriro, A., Akhtar, Z., Ricci, F. and Luccio, F. L.: Wearable Wisdom: A Bi-Modal Behavioral Biometric Scheme for Smartwatch User Authentication, *IEEE Access*, Vol. 12, pp. 61221–61234 (2024).

[12] Okawa, R., Kobayashi, R. and Yamaguchi, R. S.: Behavioral Authentication Method Focusing on the Tendency of the Days of the Week, *2022 International Symposium on Information Theory and Its Applications (ISITA)*, IEEE, pp. 224–228 (2022).

[13] Alrawili, R., AlQahtani, A. A. S. and Khan, M. K.: Comprehensive survey: Biometric user authentication application, evaluation, and discussion, *Computers and Electrical Engineering*, Vol. 119, p. 109485 (2024).

[14] Thomas, B.: Sensorlog, Available at: `https://sensorlog.berndthomas.net/` (2016).

[15] Géron, A.: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Inc. (2022).

[16] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T. et al.: Xgboost: extreme gradient boosting, *R package version 0.4-2*, Vol. 1, No. 4, pp. 1–4 (2015).

[17] scikit-learn Developers: sklearn.model_selection.TimeSeriesSplit Documentation, `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html`.