Write to Know: On the Feasibility of Wrist Motion based User-Authentication from Handwriting

Raveen Wijewickrama
University of Texas at San Antonio
raveen.wijewickrama@utsa.edu

Anindya Maiti
University of Oklahoma
am@ou.edu

Murtuza Jadliwala
University of Texas at San Antonio
murtuza.jadliwala@utsa.edu

ABSTRACT

The popularity of smart wrist wearable technology (e.g., smartwatches) has rejuvenated the exploration of dynamic biometric-based authentication techniques that employ sensor data from these devices. Despite the progress demonstrated by the scientific community, research in this area has not successfully transitioned to practice, and we are yet to see a mainstream user-authentication product based on a dynamic biometric such as handwriting/hand gestures captured using commercial wrist wearables. This work undertakes an investigative analysis to further explore why that is the case. We accomplish this by studying the feasibility and practical deployability of handwriting-based authentication techniques in the literature that utilize motion sensors on-board wrist wearables. We conduct this analysis by replicating four state-of-the-art and representative handwriting-based authentication schemes that employ wrist motion data, in order to test their viability in realistic hand-writing/gesture scenarios. By using data collected from actual human subjects in an unconstrained fashion, we comparatively evaluate the performance of these schemes with well-defined usability and security metrics. Our experimental results show that some of the tested schemes perform considerably well in practice, and are promising. However, they do suffer from several practical user-dependent and technique-specific challenges that act as roadblocks towards their wide-scale adoption in mainstream applications.

CCS CONCEPTS

• **Security and privacy → Biometrics**; • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **Activity centered design**.

ACM Reference Format:
Raveen Wijewickrama, Anindya Maiti, and Murtuza Jadliwala. 2021. Write to Know: On the Feasibility of Wrist Motion based User-Authentication from Handwriting. In *Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '21), June 28–July 2, 2021, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3448300.3468290

1 INTRODUCTION

The popularity of wrist wearables such as smartwatches has soared over the past few years. In addition to being a fashionable accessory, a large set of integrated on-board sensors enables wrist wearables to offer a wide-variety of applications besides timekeeping. This includes simple applications such as step counting and sleep time monitoring, to much more complex tasks such as continuous activity recognition [35] and personalized health monitoring [15].

More recently, the research community has been strongly pursuing the idea of employing wrist wearables, and the diverse set of on-board sensors, for (continuous) user identification and authentication tasks [5, 6, 8, 13–15, 20–22, 25, 29, 32, 36, 39]. In addition to a rich set of available sensor modalities, its position on a user's wrist makes these smart wrist wearables alluring for user authentication type of tasks, especially those based on dynamic biometrics. This is because wrists are actively used in carrying out a variety of day-to-day tasks and the associated wrist movements are unique from person to person due to distinct physiological and kinesiological differences. As a result, an analysis of wrist movements using data from wrist wearable motion sensors (e.g., accelerometer and gyroscope) can provide insight into the various user behavior or activity based authentication modalities. Some recent research efforts in this direction include gait-based authentication [6, 32], authentication based on touch input characteristics [8, 29, 36], gesture-based authentication [21, 22, 39], and handwriting-based authentication [5, 13–15, 20, 25].

This work specifically focuses on handwriting motion-based authentication mechanisms that are enabled using modern wrist wearable devices equipped with high precision motion sensors. The presence of unique characteristics in a person's handwriting along with the presence of unique wrist movements have made handwriting-based authentication using wrist-wearables, a useful application of wrist-wearable technology to investigate. Besides serving as a convenient primary or secondary (continuous) authentication mechanism for personal use (or for authenticating other devices), such a handwriting motion-based authentication technique could also serve other useful purposes. State-of-the-art research efforts in the literature [5, 13, 14, 20, 25] that employ handwriting-related wrist motion for authentication have not only investigated different practical handwriting scenarios, such as in-air handwriting and writing on a paper using a pen/pencil, but have also demonstrated compelling performance and accuracy results for successful user authentication using this modality. Furthermore, most of these efforts utilized consumer-grade wrist wearables, making them practicable and easily adoptable for real-world applications. Despite these favorable outcomes, research in this area has not successfully transitioned to practice. We are yet to see a successful mainstream mobile/wrist wearable application for handwriting-based user authentication, either as a primary or a secondary/continuous authentication factor. This begs the following questions: *what is preventing these state-of-the-art handwriting-based authentication frameworks*

WiSec '21, June 28–July 2, 2021, Abu Dhabi, United Arab Emirates

ACM ISBN 978-1-4503-8349-3/21/06…$15.00
https://doi.org/10.1145/3448300.3468290

from being adopted in mainstream mobile applications and services? Is that because they do not perform well outside of the controlled operating conditions and experimental settings used in their evaluations? And even if they do perform generally well, is their performance at a level required for being successful as a robust user authentication scheme in the wild, similar to other popular (static) biometric-based approaches (e.g., fingerprint)? Answering these questions is critical for understanding the reasons behind the lack of success and/or adoption of handwriting-related motion as a modality for user authentication in mainstream mobile and wearable applications, and for understanding how/if these challenges can be overcome.

Accordingly, we conduct an investigative analysis in this paper to answer the above feasibility related concerns, and discuss the practicality of handwriting-based authentication using wrist wearable motion sensor data. We select four state-of-the-art representative handwriting-based user authentication frameworks from the literature, and comprehensively analyze all of them using uncontrolled and unconstrained handwriting data collected from a diverse set of human subject participants in a variety of realistic writing settings. We select these frameworks due to their varying, yet representative approaches both in the design of the respective frameworks as well as the handwriting modalities. Specifically, one of the frameworks utilize carefully engineered features in both time and frequency domains along with traditional machine learning techniques such as SVM [13], while another framework only uses a basic set of temporal domain features along with a Naive Bayes classifier [5]. The third framework does not extract any features, and instead leverages on deep neural networks (DNN) [14]. Lastly, the fourth framework uses frequency domain features with a Logistic Regression model [20]. An additional insight here is that it is not appropriate to directly compare the evaluation results reported by these research efforts, primarily because the data for those experiments were collected from different groups of users (human subjects), under different conditions and setups. For an equitable comparison, we need to evaluate these schemes with data from the same group of subjects under the same experimental conditions, which we accomplish in this work. More specifically, comprehensively investigate the performance of the above mentioned four authentication frameworks against a variety of experimental conditions and parameters, including, authentication window size, training data size, writing settings (e.g., pencil, finger and air writing), robustness under environmental noise, and their ability to perform in true free-form writing.

2 RELATED WORK

User authentication has been an extensively researched topic in the literature, with a diverse body of contributions. However, here we only outline research efforts related to unique user movement-based authentication – a form of dynamic biometric – captured by means of motion sensors on a variety of mobile/wearable devices, as it is more relevant to the proposed work. A more comprehensive survey on user authentication, which includes other static and dynamic modalities and biometrics, can be found in [33, 34].

Authentication techniques that employ motion sensor data from mobile/wearable devices to construct unique user movement related biometrics have been extensively studied in the research literature [5, 6, 8, 13–15, 20–22, 25, 29, 32, 36, 39]. These efforts have either employed commercial off-the-shelf (COTS) mobile/wearable devices, including smartphones, smartwatches and smart rings, or other types of specialized devices (e.g., smart pens and hand gloves) equipped with motion sensors such as accelerometers and gyroscopes. Some of the early schemes used unique gestures made by users while holding a smartphone or while wearing a smart wrist wearable for user identification and authentication, where the gesture-related hand/wrist motion was captured by the mobile device's accelerometer and gyroscope sensors [21, 22, 39]. Other forms of contextual body movements such as users' natural gait (walking) based motion [9, 27], motion/orientation corresponding to how users hold their phone (in their hand/s) [11, 36], motion corresponding to how users answer a phone call [30] and fine-grained hand movements such as taps or typing [3, 7] on the phone captured using COTS mobile device motion sensors, have also been used to design dynamic biometrics for user authentication.

An advantage of employing such modalities (e.g., tapping, typing, etc.) as a biometric is that it can be used to continuously authenticate users in a real-time fashion. Hand movements observed during handwriting is another suitable modality for such continuous user authentication, and it has also received significant attention. We first outline authentication schemes in the literature that have employed specialized, non-standard motion-capture devices to capture handwriting related wrist/hand motions. For instance, Bashir et al. [2] proposed an authentication scheme by using a smart pen device to capture the accelerometer time-series data corresponding to two different types of handwriting scenarios: (i) writing in the air, and (ii) writing on a paper. In another research, Lu et al. [25] used a custom glove with built-in motion sensors located on the fingertip to capture in-air handwriting movements. In addition to the fact that it required a custom data collection hardware to operate, this scheme only considered writing scenarios in which users wrote a unique passcode/PIN for every authentication attempt. This severely limits its practicality, especially for continuous authentication.

In the direction of handwriting motion based authentication schemes that employ commercially available wrist wearables, Buriro et al. [4] proposed a framework similar to [25], but by using a consumer-grade smartwatch instead of a custom glove, and the users sign their name in the air to authenticate themselves. Huang et al. [16] proposed an authentication scheme using a smartwatch gyroscope in which they evaluated three gestures written/drawn in the air, namely, a star, a number eight and a triangle. A significant drawback of their approach is that it has been shown to work only for a fixed set of hand signs, and it is not evident whether their framework can be extended or generalized to normal human handwriting or signatures. As before, this limits its applicability in continuous authentication scenarios.

The following four schemes [5, 13, 14, 20] in the literature attempt to overcome the two major shortcomings of earlier handwriting motion based authentication schemes: (i) reliance on specialized motion sensing hardware, and (ii) employing fixed writing patterns or symbols for authentication. Given the diverse set of handwriting scenarios and settings considered by these schemes, and the use of COTS data collection hardware (e.g., smartwatches) instead of specialized hardware, makes them the most promising candidates for adoption by users in a practical and continuous handwriting-based user authentication application. This is also the main reason why

we shortlist them as prime candidates for a mainstream continuous authentication application in this category, and perform a rigorous comparative performance analysis of them in realistic usage settings. More detailed descriptions for each of them are outlined later in Section 4.

3 RESEARCH GOALS

We organize our research goal - studying the practical feasibility of state-of-the-art handwriting-based user authentication schemes that employ motion sensor data from wrist wearables - as a set of four targeted research questions (RQ1 through RQ4). These questions focus on studying the performance of these schemes under varying writing styles and modalities, ambient conditions and written content, and collectively provide a comprehensive performance analysis. We fix the experimental setup and evaluation parameters (Section 5), including the employed performance metrics, such that it enables us to obtain insightful answers to these questions.

- *RQ1 – How does handwriting-based user authentication using wrist motion data perform in real, unconstrained writing scenarios?*

Ideally, an effective user authentication scheme should not place any unreasonable constraints on how users should write for the scheme to work effectively. Any undue constraints or changes to the users' usual writing habits or styles will result in low usability and adoption of the scheme. This research question aims to investigate the constraints or restrictions a user authentication scheme (under investigation) places on users' writing style and how does it perform when those constraints are dropped and users are allowed to interact with the scheme using their own everyday writing style. We accomplish this by collecting handwriting-related data in a fully unconstrained setting, where participants are allowed to write in their usual habit or style. We analyze the performance of the schemes under investigation using this unconstrained writing data, and also comparatively evaluate them based on the amount (length) of writing data required for enrollment and authentication.

- *RQ2 – How does handwriting-based user authentication using wrist motion data perform for different writing modalities?*

Depending on a user's context (time, location, etc.) and the writing instrument they are interacting with, the activity of writing can assume different modalities, for example, writing on a paper with a pen, writing on a smart tablet screen using a finger, and gesture writing in the air. Ideally, an effective authentication scheme should work across multiple writing modalities, otherwise, its applicability is restricted to a limited set of user-contexts. This also limits the applicability of the scheme for continuous authentication. Thus, this research question aims to investigate how the schemes (under investigation) perform for a variety of commonly observed writing modalities. We accomplish this by collecting data for a variety of writing modalities in a fully unconstrained setting, such as, traditional pen/pencil writing on a paper, finger writing on a touchpad screen and gesture writing, and analyzing the performance of the schemes across these modalities.

- *RQ3 – How does handwriting-based user authentication using wrist motion data perform under different types of ambient noise?*

Accelerometers and gyroscopes on modern mobile devices are highly sensitive hardware/software systems. An advantage of high sensitivity is that these sensors can sample small (imperceptible) changes in motion, however, a related disadvantage is that the sampled data from these sensors is easily impacted by ambient noise. An important characteristic of an effective and practical user authentication scheme is robustness against noisy inputs. A robust (against noise) authentication system typically produces very low false negatives (i.e., has high recall), resulting in higher usability and adoption among users. This research question aims to investigate how the user authentication schemes (under investigation) perform when the sampled test motion data (or authentication requests) is noisy. We accomplish this by introducing noise in our collected handwriting related motion data, and analyzing the performance of each of the four authentication schemes on this noisy data. For this analysis, we consider different types and sources of noise commonly encountered during writing, such as writing on a table with a vibrating smartphone and writing while inside a moving vehicle.

- *RQ4 – How does handwriting-based user authentication using wrist motion data perform under user-dependent, free-form writing?*

In RQ1, we analyzed the performance of handwriting motion based user authentication by removing constraints on *how* users should write. In the same vein, in this research question, we investigate if constraints or restrictions on *what* the users write impact the performance of the schemes under investigation. We accomplish this by evaluating these schemes on free-form handwriting data collected from participants in a fully unconstrained setting, i.e., motion data corresponding to handwriting consisting of both upper-case and lower-case letters (and not limited only to specific words with specific lengths or specific letter cases). The main motivation of not restricting users to write a specific text for authentication is that it limits the applicability of the scheme in continuous authentication scenarios, since continuous authentication is about being able to passively authenticate based on whatever is being written by the user and not asking them to actively authenticate by writing specific keywords. Further, the handwritten letter/word/text cannot not be equated to a set password as nothing conceals handwritten text from onlookers, unlike passwords typed on a PC that are by default concealed on the screen. Ideally, a handwriting-based authentication scheme should be agnostic of the letter/word/text written at any given instance of the authentication attempt.

4 EXPERIMENTAL SETUP

In this section, we first outline a system model (Figure 1) that generically describes the authentication framework of all the four handwriting-based authentication schemes that we plan to comparatively evaluate in this work. Following that, we provide specific modeling and implementation details for each of the four schemes. Then, we provide details of our data collection procedure and the metrics and benchmarks used in our analysis.

4.1 System Model

In any handwriting-based authentication scheme that employs motion data from wrist wearables, users would first go through an enrollment phase in which they supply training data to the system by performing handwriting tasks while wearing a wrist wearable device on their writing hand. The motion sensor data collected from the wrist wearable during the enrollment procedure is then used to build a unique profile for the user, which is later tested against when an authentication attempt is made. This "unique

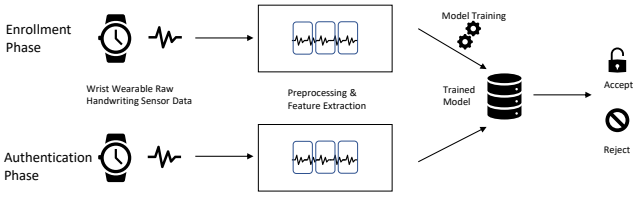


Figure 1: A generic system model for handwriting motion based authentication schemes.

profile" for each user typically takes the form of a classification function that is trained using the target user's data. The raw motion data collected during the enrollment phase is typically first pre-processed to remove/reduce noise and then utilized for the feature extraction task. The extracted feature set for the authentic user(s) is then used to train a classification model, sometimes along with labeled non-authentic user data to prevent over-fitting. As such, the classification model is a binary classifier where it will attempt to classify an authentication attempt as either authentic (valid users who enrolled with the system) or non-authentic (anyone else). In order to authenticate with the system, an enrolled user is required to perform a handwriting task which will then be compared with the user's profile, and an authentication decision is made by the system based on the output of the trained classification model.

4.2 Implementations

We implemented the four representative handwriting-based authentication frameworks (summarized in Table 1) using Python 3.6, while utilizing scikit-learn [28] and TensorFlow [1] with Keras for the corresponding machine learning and deep learning-based models employed by those frameworks. Individual implementation details of each framework's classification model are described next, labeled from M01 to M04. These implementations very closely follow the original works, and reuse their code whenever available.

4.2.1 ***M01*** *[13].* In this framework, the raw sensor data is first pre-processed by applying a low pass filter to eliminate outliers. The continuous motion data stream is then divided into smaller windows of data which are then used for feature extraction across four feature categories. First, each of the sensor axes x, y and z, along with their magnitudes, constitute as the first feature category. The second feature category includes Fast Fourier Transform (FFT), Power Spectral Density (PSD), and squared magnitude or power of FFT coefficients. The third feature category includes Discrete Cosine Transform (DCT), Discrete Sine Transform (DST), Real-Valued Fast Fourier Transform (RFFT), eigenvectors and gradients calculated for each of the x, y and z axes. The fourth and final feature category is pitch and roll [38]. Different statistics such as mean, standard deviation, variance, computed on values in these four feature categories are then combined to create an intermediate feature vector, which is further normalized and selectively shortened before being employed in the classification model training/testing tasks, as explained next. The intermediate feature set comprising of 182 features obtained in this fashion is then normalized and ReliefF [19] feature selection scheme is applied to select top-30 features from each of the accelerometer and gyroscope data, resulting in 60 features in the final feature vector. A Support Vector Machine (SVM) binary classifier is then trained for each user using this final feature vector with GridSearch based hyper-parameter tuning. The model is trained using a part of the authentic user's data and data from half the number of other users in the dataset. In the model testing phase, the remaining part of the authentic user's data and the data from the other unseen users are used.

4.2.2 ***M02*** *[5].* Instead of a sliding window approach, the M02 approach uses the entire recorded activity signal for feature extraction. In contrast to M02 [5], which employed short and specific transcription tasks during data collection, our common data collection process (outlined in 4.3) involved long handwriting tasks in which users wrote continuously for few minutes. Thus, we utilize a sliding window approach to divide the raw signal (obtained during our data collection) into multiple smaller windows which are then independently used for feature extraction. For each window, the mean, standard deviation and average absolute difference for all 3 axes are calculated and used as features. The peak positive value and the peak negative value are also extracted from each axis and used as features. The average resultant acceleration, which is the mean of square root of sum of the squared values of x, y and z axes, is also computed and used as a feature. A combination of these features results in a vector of 16 features for each sensor type and a final feature set of 32 features. A Naive Bayes model is then trained and tested in a per-participant fashion, similar to M01.

4.2.3 ***M03*** *[14].* The M03 approach employs a Deep Neural Network (DNN) which is created using two consecutive 1D convolutional block layers, followed by a bi-directional LSTM layer for each sensor axis, which are then concatenated and fed into a fully connected layer. Each of the convolutional blocks consists of two 1D convolutional layers with 32 and 64 filters, respectively, with a kernel size of 3 and a "relu" activation followed by a batch normalization layer, a max pooling layer of size 2 and a dropout layer with a 0.5 dropout rate. The output of each convolutional block (per axis) is then sent through a bidirectional LSTM layer with 10 neurons. Each of the output from each axis/convolutional block is then concatenated, sent through another layer of dropout with a rate of 0.5 before finally being input to a dense or fully connected layer with 2 neurons. The dense layer consists of 2 neurons which represents the number of classes in the classification problem along with a softmax activation. The model is fitted using Categorical Cross Entropy loss with Adam [18] optimizer. As DNNs require large amounts of data to train an effective model, M03 [14] uses a data augmentation step in which synthetic data is created using the existing data, in order to increase the training data size. In our experiments, we implement the same data augmentation step where random windows of data are selected to generate new synthetic data, which is repeated till the dataset becomes 4 times the size of the original set [37].

4.2.4 ***M04*** *[20].* As the original work in M04 is designed for signature verification, entire motion data corresponding to each signature sample is taken as an input for feature extraction. In order to adapt it to a handwriting authentication scenario, we extract small windows from the data similar to the previously described frameworks. The windows are then normalized and transformed into the frequency domain by Discrete Cosine Transform (DCT). The first 20 DCT coefficients are then extracted and fed as input into the feature extraction phase. In feature extraction, a set of authentic user data windows are selected as templates for the authentic

Table 1: Overview of the four representative authentication frameworks used in our study.

	Description	Device	Sensors & Frequency	Features	Techniques
M01 [13]	Handwriting authentication using paper and pencil as writing tools	Android Smartwatch	Accelerometer, Gyroscope at 100Hz	Both temporal and frequency domain features	SVM, MLP
M02 [5]	Handwriting authentication using paper and pencil as writing tools	LG G Watch	Accelerometer, Gyroscope at 20Hz	Only temporal domain features	NB, MLP
M03 [14]	Handwriting authentication using paper and pencil as writing tools	LG Urbane 2 Smartwatch	Accelerometer at 100Hz	No specific features extracted. Uses raw sensor data.	DNN
M04 [20]	Signature Verification using stylus tablet as writing tools	Microsoft Band	Accelerometer, Gyroscope at 62Hz	DTW distance based features in frequency domain	Logistic Regression

user. The remaining authentic user data was then split into training and testing sets. The training set of an authentic user and half the number of non-authentic users are selected for model training. For each of the sample windows in the training set, the DTW score between each window and the authentic user template windows are computed. The DTW score is calculated individually for each axis of a query sample and the templates in the template set, the lowest score obtained between a template and a query sample is then added to the feature vector (the feature vector consists of 6 features, one for each axis of each of the sensor). The training data created in this fashion is then used to train a Logistic Regression based classification model.

4.3 Data Collection

We recruited a demographically diverse set of 21 participants for data collection in three popular writing scenarios, namely *pencil-writing*, *finger-writing* and *air-writing*, utilizing seven participants per scenario. The participants were recruited via advertising flyers around the university and their ages range between 18 to 30 (σ = 4). The participants were either given a Sony Smartwatch 3 or a LG Watch Urbane to wear on their writing hand (right hand) for the entire experiment. A Samsung GT-N5110 Android tablet was used to display the writing content. An Android Wear app which records accelerometer and gyroscope data at 200 Hz from the smartwatch was developed along with an app for the tablet to display the writing content. In the pencil writing experiment, the participants were provided with a pencil and paper and a table to write on. They were also given a height-adjustable chair, which they could adjust based on their individual/personalized comfort and writing positions. The finger-writing task was carried out in a similar way, except that the participants were given an area on the screen of the android tablet for the writing task. In the air-writing setting, the participants were given a chair with no obstructions for the free movement of their writing hand. In all scenarios, the participants were asked to stick to their natural handwriting styles and pace, and no time limit was given to complete the tasks. The participants were asked to write the content displayed on the tablet on the corresponding writing surface. For example, if the tablet displays the word "test", then a pencil-writing participant would write the word "test" on the paper given to them using a pencil. In each of the writing scenarios, each participant wrote English alphabets displayed in random order. Each alphabet was written 10 times, totaling to 260 alphabets for lowercase and 260 for uppercase alphabets followed by 40 words. These data collection procedures were approved by our university's Institutional Review Board (IRB).

We also evaluate robustness using a setup where the attacker tries to falsely authenticate by mimicking the handwriting of an authentic user. For this mimicking/impostor attack evaluation, 5 participants (potential victims) were recruited. Their handwritten text of six random Harvard sentences [31] and a video of their writing (which includes their hand/wrist movements and wrist positioning) were recorded.

An attacker was then asked to observe the victims' text and video for practice, i.e. the paper with the handwritten text of the victim or the screen recording of the text written on the tablet surface for pencil-writing and finger-writing scenarios respectively. In addition, the attacker also carefully watched/observed the video taken of the victim's handwriting before performing the final impostor attack. We only considered the finger writing and pencil writing settings for this analysis because air writing does not produce physically recorded written text that can be reviewed by an attacker for mimicking purposes.

4.4 Benchmarks and Evaluation Metrics

We comparatively evaluate the authentication frameworks using the widely accepted metric of Equal Error Rate (EER), which is the point at which the False Rejection Rate (FRR) equals the False Acceptance Rate (FAR) [12]. FAR is the probability an authentication system incorrectly authenticates an unauthorized user or impostor as an authorized user. On the other hand, FRR is the probability that an authentication system incorrectly rejects an authorized user as unauthorized (failed authentication). These metrics depend on the decision threshold of the classification system. A strict decision threshold would imply lower number of false positives, i.e. the probability of an unauthorized user or impostor getting authenticated is lower, but actual users may get rejected for slight anomalies in their writing with higher probability. On the other hand, with a relaxed threshold, there is a higher probability of an unauthorized user or impostor getting authenticated and the probability of authorized users getting rejected for slight anomalies in their writing would be lower [24]. This trade-off between FAR and FRR is application specific. A high security application in which an entry of an unauthorized user is disastrous must require a very low FAR regardless of the possible inconvenience that authorized users may experience [17]. For reporting comparative analysis on the framework performance, we use EER as the primary metric, and employ the FAR and FRR metrics wherever relevant. In general, a lower EER value is a good indicator of a balanced and robust authentication framework.

5 PERFORMANCE EVALUATION

All the four authentication frameworks are evaluated per participant, by labeling the target participant's data as authentic and all other participants as non-authentic or impostor when training and testing the models. The overall evaluation results are then averaged

across all the target participants. We present EER values for comparison between the handwriting-based authentication schemes, utilizing realistic data collection procedure outlined in Section 4.3. We conduct our comparative performance evaluation with respect to varying experimental parameters, as categorized next.

5.1 The Effect of Authentication Window Size

The window size parameter is intuitively the continuous window of time a user needs to write in order for the authentication framework to effectively perform its functionality. Or in other words, it is the number of continuous motion sensor data samples of handwriting activity used by the authentication framework for the classification task. The window size is a vital experimental parameter because a smaller window may not contain sufficient information for successful authentication, while a larger window size will require a longer time to process and for the authentication task to complete. If an authentication system requires users to write for long periods of time for every authentication attempt, it can result in usability issues as well. In our experiments, we test window sizes ranging from 15 to 60 seconds for both the enrollment and the authentication phases, and we present the results next.

In the M01 window size analysis shown in Figure 2a, we see that while all the window sizes (15, 30, 45 and 60 seconds) produce a mean EER of less than 0.10, a window size of 60 seconds performs slightly better (lower EER) for both pencil and finger-writing scenarios. In the air-writing scenario, the highest EER is recorded for the 15 second window at 0.16 (σ = 0.17) and the lowest/best EER is recorded again for the 60 second window which is 0.10 (σ = 0.12). Due to the additional freedom when writing in the air, we observe that users tend to write larger characters which in turn results in lower number of characters per window. Thus, for the air-writing scenario, a higher window would capture more characters and more discriminative features across users. Although the difference between the obtained EER values across different window sizes is insignificant in pencil and finger-writing scenarios, in the air-writing scenario 60 second windows have a clear advantage over the other window sizes with the next lower window size of 45 seconds recording an EER of 0.14. The trends are similar to the results obtained by the authors of M01 at a mean EER of 0.11, especially in the pencil and finger-writing scenarios.

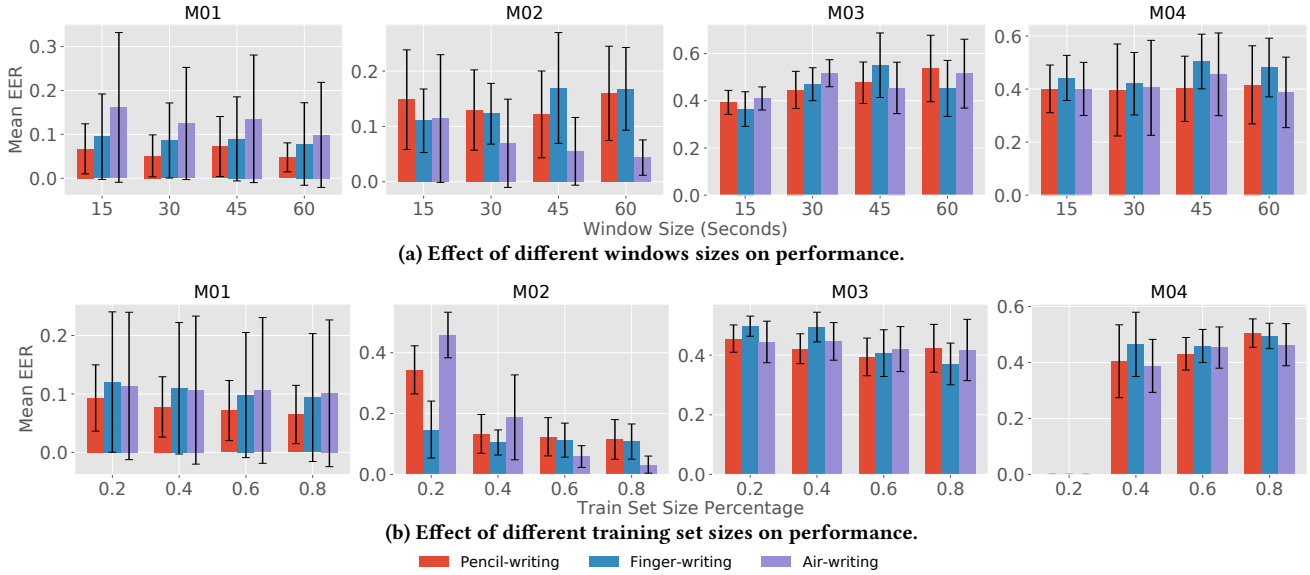
In M02 window analysis (Figure 2a) for the pencil-writing scenario, we observe that the best mean EER is 0.12 (σ = 0.08) with a window size of 45 seconds. Although the performance difference between window sizes 30 and 45 was low, at 60 second windows the performance deteriorates to an EER of 0.16. In finger-writing scenario, we observe that the best mean EER is 0.11 (σ = 0.05) at 15 second windows, and it can also be seen that the performance worsens with the size of the window. A possible explanation for the degrading performance at larger window sizes could be that larger window sizes lead to lower number of training examples, which results in an underfitted classification model. In the air-writing scenario of M02, we observe that the best mean EER is 0.04 (σ = 0.03) at 60 second windows and 45 second windows is second best at 0.05 (σ = 0.06).

As shown in Figure 2a, in all 3 writing scenarios M03 produced the lowest EER values when a 15 second window size is used compared to other window sizes. Specifically, the resulting EER values are 0.39 (σ = 0.05) for pencil-writing, 0.36 (σ = 0.07) for finger-writing and 0.4 (σ = 0.04) for air-writing scenarios. In contrast, the experimental results of M03 as reported in the original publication shows that it was able to achieve slightly better results with a mean EER of around 0.3 compared to the results we have achieved. One possible reason why the DNN model of M03 performs better at 15 second windows could be because at a lower window size there would be more training examples which is advantageous when training a DNN model. In the original proposal, a voting based multiple window fusion technique is used after the DNN classifier step to lower the EER to 0.07, but we did not observe significant performance improvement even after such a multi-window voting mechanism. In M04, the mean EERs across all window sizes for pencil-writing scenario is around 0.4. In the finger-writing scenario for M04, the lowest mean EER for a window size of 30 seconds is around 0.42 (σ = 0.12). Similar to M01 and M02 air-writing scenarios, M04 air-writing scenario has the best EER at a window size of 60 seconds which is 0.39 (σ = 0.13).

In summary, M01 performs the best (EER of less than 0.1) across all schemes in 15 seconds window sizes, while M02 shows the best performance (with an EER of 0.12) for air-writing at the same window size. Thus, from a practical perspective, our results show that M01 is more desirable due to its comparatively better performance at lower window sizes, especially for the popular pencil and finger-writing scenarios.

5.2 The Effect of Training Set Sizes

We next analyze the effect of training set sizes on the performance of the authentication frameworks M01-M04. To do so, we test each framework with training set sizes between 20% to 80% of the total available user (participant) data. However, the total amount of time and training examples for each user varies due to varying writing speeds. One user may have more characters or words written during a particular window of time compared to another. This also highlights that these authentication frameworks may require varying amounts of data during the user enrollment phase, depending on the targeted accuracy or error-rate thresholds. Requiring larger amounts of training data to achieve a better performing framework is not very convenient from the end-user perspective, thus making the mainstream adoption of such schemes difficult. As can be seen from our experimental results in Figure 2b, M01 does not have any significant performance improvement above the training set size threshold of 40% for any of the writing scenarios. Similarly, M02 (Figure 2b) also does not have any significant performance improvement above the training set size threshold of 40% for the pencil-writing and finger-writing scenarios. However, in the air-writing scenario, a considerable performance improvement (for M02) is observed when the training set size is set above 40% of the total user data. Specifically, the EER at 40% training set size is 0.19 (σ = 0.07), and when the training set size is set to 60% it dropped down significantly to 0.06 (σ = 0.04), which further drops to 0.03 (σ = 0.03) when using 80% of the data for training. This again indicates that due to the extra freedom and the higher time (larger characters) that occurs during air-writing, the framework hugely benefits from a having higher amount of training data allowing it to generalize well on the test data.



(a) Effect of different windows sizes on performance.

(b) Effect of different training set sizes on performance.

Figure 2: Performance based on varying window sizes and train set sizes.

In the evaluation of M03 and M04 (Figure 2b), we observe that both these frameworks still produce mean EERs of over 0.35 across all writing scenarios even with higher training set sizes (The training set size of 20% is not evaluated for M04 as it had insufficient examples for constructing both the required template and training sets). Although M03 with its DNN model could benefit from a larger set of training data, the performance increases observed were minimal, again indicating the need for high training data volumes. The overall poor performance of M04 on the other hand could be indicative of the fact that the specific features that were used by the scheme did not generalize well for the continuous handwriting based authentication task.

In summary, framework M01 and M02 perform considerably better at even lower train set sizes indicating that these frameworks do not require large amounts of training data for successfully operating as an authentication scheme.

5.3 Different Writing Settings

Next, we present a detailed evaluation for each authentication framework in each of the three writing scenarios (Table 2).

5.3.1 Pencil-writing. In the pencil-writing scenario, our experiments with M01, which uses a SVM binary classifier, demonstrated a mean EER of 0.05 across all participants (σ = 0.03). The lowest EER we observe for a single participant is 0.01 and the worst EER for a participant is 0.14. Furthermore, 6 out of 7 participants show a mean EER of less than 0.1. M02, with a Naive Bayes classifier, has a mean EER of 0.15 (σ = 0.10) across all participants for the pencil-writing scenario. We also observe that for 5 out of the 7 participants, the EER is below 0.20. The best M02 EER for a participant is 0.02, while the worst is 0.32. For M03, the mean EER across all participants is 0.39 (σ = 0.05). The lowest recorded EER for a participant is 0.34 and the highest is 0.47. The best mean EER of framework M04 is 0.40 (σ = 0.09) for a 30 second window. The lowest EER recorded for a participant in M04 is only 0.34 and the highest is as high as 0.52 in M04. Moreover, only one participant has an EER below 0.40 in M04. In summary, for the pencil-writing scenario, we observe that M01 produced the lowest EER compared to all other frameworks.

5.3.2 Finger-writing. In the finger-writing scenario, M01 has the best EER (compared to all other schemes) of 0.08 (σ = 0.09), with 6 out of the 7 participants having less than 0.12. The framework M02 shows the second best performance with a mean EER of 0.11 (σ = 0.06) and a best EER of 0.03. M03 has a slightly higher mean EER of 0.36 for finger-writing, in which two of the participants have a mean EER of 0.26 while all other participants' EERs are over 0.40. Lastly, M04 has the worst performance for finger-writing with a mean EER of 0.42 (σ = 0.12) and with only one participant showing an EER below 0.3. Similar to the pencil-writing scenario, M01 produces the best performance (lowest EER) for finger-writing scenario.

5.3.3 Air-writing. In the air-writing scenario, M01 has a mean EER of 0.10 (σ = 0.12) with only one out of the 7 participants recording an EER higher than 0.10. The M02 framework demonstrates the best performance overall out of all the frameworks for air-writing with a mean EER of 0.08 (σ = 0.05), with all except one participant showing an EER of over 0.06. The M03 framework has a mean EER of 0.4 (σ = 0.08) with only one participant showing an EER of below 0.3, while M04 shows similar performance with a mean EER of 0.39 (σ = 0.06). In summary, while M01 demonstrates the best performance (lowest EER) for pencil and finger-writing scenarios, M02 has the lowest EER for the case of air-writing.

5.4 The Effect of Sampling Frequency

Even though modern wrist wearables feature highly precise motion sensors, sampling these sensors at a high rate is an energy intensive operation, which significantly impacts the device's battery life, and thus, treated as a critical design factor. Lu et al. [23] demonstrate that higher sampling frequencies of motion sensor data results in considerably higher battery power consumption. Specifically, they

Table 2: Performance summary (measured in EER).

	M01	M02	M03	M04
Pencil-Writing	0.05	0.15	0.39	0.40
Finger-Writing	0.08	0.11	0.36	0.42
Air-Writing	0.10	0.08	0.40	0.39

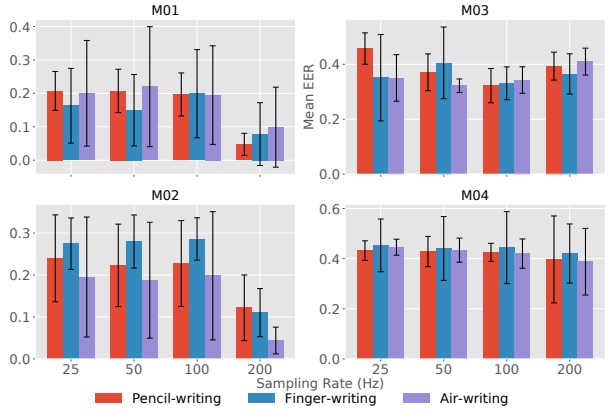


Figure 3: Effect of sensor sampling rates on performance.

show that a rate of 200 Hz consumes 6.3% (per hour) of the battery power of the device on average (tested on a Sony Smartwatch 3) and battery power consumption reduces when the sampling rate is lowered with only 4.4% (per hour) at a 100 Hz sampling rate and 2.0% (per hour) at a 50 Hz sampling rate. In other words, handwriting authentication frameworks that require fine-grained motion data (i.e., sampled at high frequencies) for performing well could adversely impact the device's battery charge (i.e., drains it faster) and requiring frequent battery recharges. This ultimately will adversely impact the usability and adoption of such schemes by end-users. Thus, in this set of experiments we evaluate how the four authentication frameworks M01-M04 perform under motion sensor data sampled at different frequencies. Schemes that perform reasonably well at lower frequencies would obviously be much more energy (battery) efficient, and preferable by end-users.

Our experiments with M03 and M04 show (Figure 3) that lower sampling rates produce comparable performance across all the writing scenarios, i.e., their performance does not vary much with change in sampling frequencies. But as seen in Figure 3, frameworks M01 and M02 produce significantly worse EERs at lower sampling rates, compared to the original sampling rate of 200 Hz. In summary, none of the analyzed frameworks produce reasonable levels of EERs at lower sampling frequencies. This highlights a significant challenge, especially, towards use of these frameworks in continuous authentication scenarios as it would require periodic sampling of motion sensors, which will in turn impact battery longevity.

5.5 The Effect of Environmental Noise

Next, we evaluate the impact of environmental noise on the performance of the four authentication frameworks M01-M04. For this, we separately record a few types of background noises that users could encounter during each of the writing scenario. We then superimpose this pre-recorded motion sensor noise over the raw handwriting related motion sensor data obtained from our study participants, prior to using the data for training and testing of the four authentication frameworks. For finger-writing and pencil-writing scenarios, the motion noise is emulated by placing a vibrating phone on the table (writing surface) at close proximity. For the air-writing scenario, the accelerometer noise emulation is obtained from a moving vehicle. After re-performing our experiments with the noisy data, we observe (Figure 4) that all frameworks demonstrate a considerable degradation of performance. We believe that our results are still relatively optimistic as we only introduced one type of noise for each of the writing scenarios. In practice, there could be a combination of various additional environmental noises polluting the device's motion sensor data which will further worsen the performance of these frameworks.

5.6 Convenience vs. Security

We also analyze how adaptable each of these frameworks are in terms of their potential target application, i.e. whether they are more suited for a high security application or for a high user-convenience application. High security applications such as access control to government intelligence, military applications or other highly sensitive data could tolerate high FRR, but FAR needs to be kept at a minimum level. Similarly, usability focused applications such as consumer-grade smartphone unlocking which prioritize user convenience, could allow a slightly higher FAR as a trade off in achieving a minimal FRR.

From our experiments with M01 (Figure 5a) we see that in the case of pencil-writing the EER occurs approximately at 0.5 decision threshold. However, a slight increase in the decision threshold would result in a sudden increase in the FRR (over 0.2), while the FAR only reduces slightly and converges around 0.8. This suggests that if a stricter (higher) decision threshold is chosen with high security in mind, the rate at which unauthorized users may mistakenly gain access decreases only slightly. However, actual users would suffer considerably as they are likely to fail authentication approximately 2 or more out of 5 attempts for decision thresholds above 0.5. Although the low EER of 0.05 suggests that M01 is generally a suitable framework for balancing convenience and security, it lacks flexibility towards adjusting the decision threshold. The M01 framework in finger and air-writing scenarios show similar characteristics when it comes to the adjustability of the decision threshold. Moreover, as the EER for these writing scenarios lie at approximately 0.10, the room for adjustability is further reduced when compared to the pencil-writing scenario. As seen in Figure 5b, for all three writing scenarios in M02, the ERR occurs at a decision threshold below 0.3. When adjusting the decision threshold in the pencil-writing scenario (EER=0.15), an increase of the threshold from 0.3 to 0.5 drops the FAR from 0.09 to 0.04. However, at the same time, FRR almost doubles from 0.15 to 0.29. A similar pattern was also observed in the finger-writing scenario. Taking a closer look at the air-writing scenario in M02, we observe that the threshold could be made stricter (up to a 0.5 threshold) in trying to achieve a much more secure system by keeping the FRR under 0.2 and reducing the FAR to less than 0.03. For the M03 framework (Figure 5c), adjusting the decision threshold either for security or for convenience would significantly increase FRR and FAR values resulting in an EER over 0.5. Very similar patterns were also observed for the M04 framework (Figure 5d). This further demonstrates that M03 and

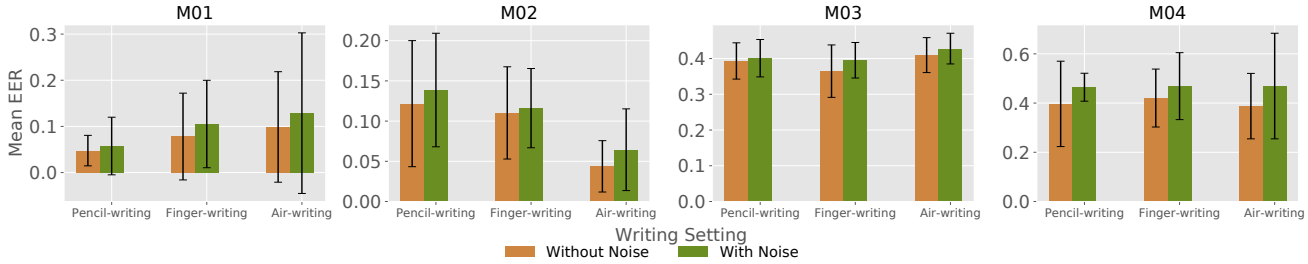


Figure 4: Effect of environmental noise on performance.

M04 are rather unsuitable for both security and usability oriented applications. In terms of adjustability of the threshold depending on the use case (security or usability), M02 demonstrates better versatility across all frameworks and all writing scenarios.

5.7 Comparison with Other Modalities

We next compare the handwriting-based authentication frameworks against other modalities of biometric authentication. The best mean EER among the four handwriting-based frameworks is for M01 in pencil-writing scenario (0.05) followed by finger-writing (0.08). Framework M02 came in second place with an EER of 0.15 for pencil-writing and 0.11 for finger-writing. M02 performed the best among all frameworks in air-writing scenario with an EER of 0.08, with M01 following behind at 0.1. Evidently, these EER values are much higher compared to the mainstream fingerprint based authentication methods (Table 3) which can be as low as 0.000022 [10, 40]. On the other hand, the handwriting-based authentication schemes' EER values are equitable to that of other motion sensor based authentication frameworks. Specifically, frameworks based on gait [26] and gesture [39] have EERs of 0.07 and 0.04, respectively. This is not surprising because, while modalities such as fingerprint and iris have matured to become reliable first factor authentication mechanisms, motion based authentication frameworks (including handwriting-based) are primarily being explored as a means for second factor or continuous authentication.

Table 3: EER attained by other authentication modalities.

Biometric	EER	Used at Mainstream Consumer Level
Fingerprint [10, 40]	0.00022	Yes
Iris [41]	0.006	Yes
Gait Based [26]	0.07	No
Gesture Based [39]	0.04	No

5.8 Mimicking Attack

To evaluate the mimicking attack, we look at the False Acceptance Rate (FAR), because a higher FAR is indicative of a more successful attack. FAR is calculated at the same decision threshold where the EER occurs. We observe an overall increase in FAR in M01 for both pencil and finger-writing scenarios from < 0.1 to over 0.2 during the attack. For M02 we observe a slight increase in FAR in pencil-writing from 0.09 to 0.11, followed by finger-writing from 0.8 to 0.15. For M03 and M04, in pencil-writing scenario, we observe an increase in FAR from 0.33 to 0.40 and from 0.26 to 0.49, respectively. But for finger-writing scenario, both frameworks showed a decrease in FAR from 0.43 to 0.35 and from 0.66 to 0.37, respectively. In summary, M01 is more vulnerable to a mimicking attack in both finger and pencil-writing scenarios followed by M02, which shows a relatively lower increase in FAR. While M03 and M04 have higher FARs even in non-mimicking attack scenarios, we observe that only the pencil-writing scenario is affected by the attack.

6 FACTORS IMPACTING PERFORMANCE

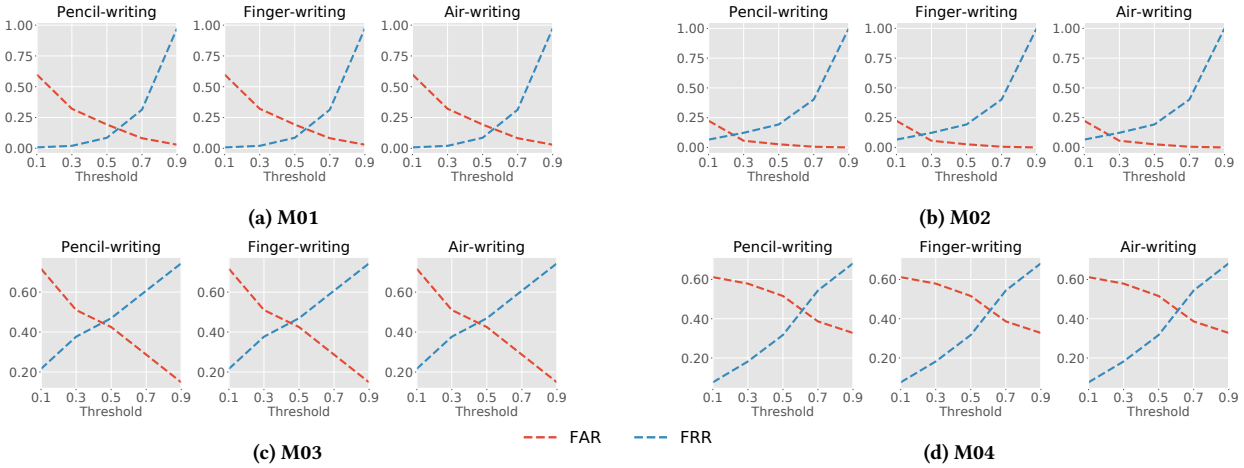
In this section, we comprehensively investigate additional factors that could potentially impact the performance of the authentication frameworks evaluated in this work. We focus on framework-specific factors, namely, feature selection and its impact on the learning-based classification models employed by all the four frameworks followed by participant-specific factors, namely, how the diversity in handwriting styles and techniques impact performance.

6.1 Feature Analysis

As all the authentication frameworks that we study in this paper employ some type of a supervised learning-based classification function, our first objective is to further investigate which features (computed from the training data) have the most impact on the framework performance, and if the performance varies significantly with a change in the feature set. We first evaluate the M01 framework, which employs both temporal and frequency domain features and ideally uses the top-30 features out of a total of 182 features calculated for each of the accelerometer and gyroscope raw data stream. In other words, it uses a total of 60 features (30 computed from the accelerometer data and 30 from the gyroscope data). We re-tested the M01 framework by reducing (choosing top-15 instead of top-30 features for each sensor stream) and increasing (choosing top-60 instead of top-30 features for each sensor stream) the size of the employed feature sets and observing its effect on the overall performance of the scheme under different writing scenarios.

Our experimental results for this analysis, outlined in Figure 6, show that for the pencil and air writing scenarios, the mean EERs obtained for the reduced feature set case (top-15 features for each sensor stream) is quite comparable to the regular case (top-30 features for each sensor stream), indicating that the framework performs fairly well even when using a lower number of features. Finger writing scenario was an exception here, where the performance for the reduced feature set case was slightly worse (mean EER around 0.11) compared to the regular case (mean EER just under 0.08). However, we observed that, as the size of the feature set increases, the performance of M01 framework worsens for all three writing scenarios. This indicates that the number of features originally selected by the M01 framework for the final feature set (i.e., top-30 from each sensor stream) provides the optimal performance.

Among these top-30 features per sensor stream, we observed that close to 50% of the features were related to general statistics, such as mean, standard deviation and variance of both time and frequency



(a) M01 (b) M02 (c) M03 (d) M04

Figure 5: False acceptance rates vs. false rejection rates at different thresholds.

domain features. However, for all the three writing scenarios for M01, we observed (across all participants) that only a very few frequency domain features were selected (in the top-30) during the feature selection step compared to the time domain features. Also, having a larger intermediate feature set (182 features) allows the model training pipeline to select the best features for a given participant. We believe that this is also one of the main reasons why M01 performed well across all writing scenarios and all participants, compared to other schemes.

In contrast to M01, the M02 framework does not include a feature selection step in the processing pipeline, and computes and uses only time domain features for model training and verification. Thus, for M02, we further investigate for each writing scenario the best set of features and study their impact on its performance. To this end, we select, top 8, 16 and 24 features out of the total 32 feature set. Our analysis shows that a top-8 feature selection step provides minor performance improvements (lower EER) across all writing scenarios. We further observed that for the pencil writing scenario, 6 out of the top-8 features are accelerometer features. This does reflect on the actual writing scenario since during pencil writing there would only be very minute angular accelerations. In finger writing, we see an equal number of features from both accelerometer and gyroscope sensors since more wrist movements could be observed during finger writing. Air writing shows similar behavior to finger writing with roughly an equal number of features from both the sensors in the top-8. Even though the M02 framework was designed for the pencil/pen on paper handwriting scenario, we observed in our evaluations that it performed rather well for the air writing scenario, compared to the pencil and finger writing scenarios. We believe that this may be because the features computed in M02 are much more responsive to significant movements of the wrist (including the arm), which is the case during air writing. To perform better for the pencil and finger writing scenarios, M02 may need to include a much more carefully computed set of features that precisely captures the subtle movements of a user's wrist.

The DNN based classification model for the M03 framework employed the raw accelerometer data stream, and thus no feature extractions were required/performed. As a result, we were not able to perform an equitable feature level analysis for M03, similar to

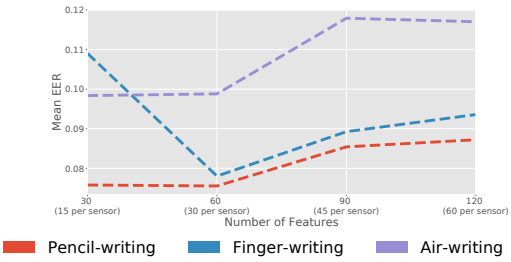


Figure 6: M01 feature analysis.

what we did for M01 and M02. However, we believe that a complex DNN-based classification function, such as the one used in M03, would require large amounts of training data to achieve acceptable levels of performance, comparable to other frameworks analyzed in this paper. This requirement of a collection of a large training dataset is not feasible during a relatively short enrollment period in practice, and is the biggest shortcoming of M03.

To recollect, framework M04 was originally proposed for signature verification. Our goal was to investigate if such a scheme could be adapted for a free-form handwriting based authentication. Our experimental results and analysis, as presented in Section 5, show that M04 features are not well-suited for authentication using motion data corresponding to free-form writing with it producing EERs around 0.4 across all the writing scenarios. We believe that one of the main reasons for this is because the written text corresponding to a signature has very little variability, compared to free-form writing which has a lot of variability (for example, the same letter can be written in different ways by the same person). Thus, features in M04 which are computed based on DTW (or similarity) scores between the query or test sample and a set of template (or training) samples works well for signature based authentication, but do not generalize well for authentication based on free-form writing.

6.2 Participant Handwriting Specific Factors

As users have different (often, unique) handwriting styles and traits, our second objective is to further investigate which participant-dependent handwriting factors significantly impact the performance of the frameworks under consideration. To this end, we first

carefully analyze the inconsistencies or irregularities that could be observed in users' handwriting traits and styles. We observe (during our data collection experiments) that there are significant irregularities in handwriting by the same participant. One key factor that contributed to this irregularity is the number of strokes a user or participant employs when writing certain characters. For example, most users write the English uppercase letter 'B' with 2 strokes, but at times the same user may write it using just one stroke (for example, when in haste or hurry) resulting in a completely different wrist motion. To further characterize this irregularity in users' writing styles/behaviors, we analyze the finger writing data which we collected using a smart tablet as the writing surface. From this data, we observe that several users employ varying number of strokes for at least 3 out of the 26 lowercase alphabets and for at least 5 out of the 26 uppercase alphabets. These character-level inconsistencies or irregularities easily propagate to words and sentences, and thus to any models trained on these prolonged writing constructs. In summary, handwriting irregularities may adversely impact the performance of authentication models trained using handwriting-related motion data, and such models may not generalize well to the different writing styles by the same user in different contexts.

Another factor that could adversely impact the performance of authentication frameworks employing handwriting related wrist motions is the positioning of the wrist while writing. During the authentication enrollment phase, a user may have positioned his/her wrist at a certain stance with respect to the writing surface, however, during the authentication (or verification) phase, that stance may be different or may have changed. The stance or positioning of the wrist while writing, together with the angle of the wrist, could significantly affect the way a user's wrist moves while writing, which inadvertently affects the performance of these authentication schemes. Additionally, the tendency of the wrist wearable device striking or coming into contact with the writing surface while writing (especially, during pencil and finger writing scenarios) could also introduce significant noise in the captured wrist motion data and affect the authentication performance. These factors and issues were very commonly observed during our handwriting data collection experiments which were done in a realistic and completely unconstrained setting.

Achieving high performance (or accuracy) when building a classifier for hand-writing motion based user authentication will require taking into account all these inconsistencies during the model training phase. But the main question that arises to this end is: *how easy or practical it is to replicate all such instances of inconsistencies (which are highly dependent on external and, at times, unpredictable factors) during the enrollment/training phase.* As an example, in a traditional fingerprint-based (static) biometric authentication system, users would simply be required to touch the fingerprint reader with different pre-defined angles/portions of the finger during the enrollment phase. But, to train a handwriting-based (dynamic) authentication system in a similar fashion, it will be non-trivial to enumerate all the pre-defined set of scenarios that users must write in during the enrollment phase. Furthermore, from a user convenience point of view, users may be reluctant to spend too much time in the enrollment phase. However, in a use case where handwriting-based authentication is used as a continuous authentication method, the classification models can be continually updated/improved over time with more user data passively collected during authentication events. In this way, the difficulty in training the system due to various participant-specific and environmental factors discussed above can be overcome to a certain degree.

7 DISCUSSION & CONCLUSION

The widespread use of smart devices in activities of everyday life, along with various apps handling users' private information, has made authentication of these devices/applications essential. Thus, the need for secure, yet convenient mechanisms for user authentication have become imperative. In this work, we evaluated four state-of-the-art handwriting-based authentication schemes with the goal of understanding the true potential and practicality of these schemes, followed by an extensive analysis of possible technical challenges faced by such authentication schemes. We comparatively analyzed these schemes against vital parameters, such as the writing sample window sizes, training data sizes, and performance at different sampling rates. Findings related to our specific research goals are summarized in Table 4. We further discussed how each of these schemes perform in terms of convenience and security, which is often a trade-off when it comes to authentication mechanisms.

When considering different writing settings, while air writing has shown comparatively better performance with an EER below 0.05 specifically with scheme M02, the practicality of using air writing for a continuous authentication scheme is questionable. It is unlikely that air writing being used as a writing mode for extensive writing tasks, since prolonged air writing could be tiring for the arm, making it unsuitable as a continuous authentication scheme. Both pencil-writing and finger-writing based authentication seem reasonably practical in real life, but finger-writing has a slight advantage since it does not require any other tools such a pencil and paper.

Prior to concluding, we would like to highlight some additional shortcomings of the handwriting motion-based authentication frameworks studied in this paper. All the authentication frameworks evaluated in this paper employ a binary classification function that needs to be trained using both authentic and non-authentic user data. Such a trained model cannot be developed purely on the user-end (e.g., a user's device), but it has to be developed on some service provider end who has access to data from multiple users. In other words, when new users want to enroll in such an authentication framework, they may have to share their personal motion data (corresponding to their handwriting) with a service provider for building a personalized authentication model for themselves. This raises significant privacy concerns for the users. Moreover, training the model on the service provider end is not efficient as any (or all) model updates (e.g., as required in the case of continuous authentication) would need to be communicated to the provider resulting in significant communication cost and latency.

Our concluding perspective on handwriting-based authentication is that while its immediate adoption is uncertain, it can show major improvements in the future as smart wearables come equipped with more precise and efficient sensors. When that occurs, handwriting-based authentication can potentially become another mainstream mechanism for user authentication.

Table 4: Summary of results.

RQs	Insights Gained
RQ1	Frameworks such as M01 and M02, performed reasonably in terms of usability/practicality under constrained writing, but required significant proportions of training data which may hinder their adoption as a mainstream authentication scheme.
RQ2	Certain frameworks (especially M01) demonstrated good potential for adaptation across different writing modalities.
RQ3	A degradation of performance was observed in the presence of ambient noise, which suggests that in real-life usage performance could further degrade raising practicality issues for mainstream adoption.
RQ4	M01 and M02 evaluations suggest that free-form handwriting-based user authentication is achievable with trade-offs between convenience and security showing potential for mainstream adoption.

ACKNOWLEDGMENT

Research reported in this publication was supported by the National Science Foundation (NSF) under award numbers 1828071 and 1943351.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, and et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.
[2] M. Bashir, G. Scharfenberg, and J. Kempf. 2011. Person authentication by handwriting in air using a biometric smart pen device. *Proceedings of the Biometrics Special Interest Group (BIOSIG)* (2011).
[3] A. Buriro, B. Crispo, F. Del Frari, and K. Wrona. 2015. Touchstroke: Smartphone user authentication based on touch-typing biometrics. In *International Conference on Image Analysis and Processing (ICIAP)*. Springer, 27–34.
[4] A. Buriro, R. Van Acker, B. Crispo, and A. Mahboob. 2018. AirSign: a gesture-based smartwatch user authentication. In *International Carnahan Conference on Security Technology (ICCST)*. 1–5.
[5] F. Ciuffo and G. M. Weiss. 2017. Smartwatch-based transcription biometrics. In *IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. 145–149.
[6] G. Cola, M. Avvenuti, A. Vecchio, G. Yang, and B. Lo. 2015. An unsupervised approach for gait-based authentication. (2015), 1–6.
[7] K. R. Corpus, R. J. DL Gonzales, A. Scott Morada, and L. A. Vea. 2016. Mobile user identification through authentication using keystroke dynamics and accelerometer biometrics. In *Proceedings of the International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. 11–12.
[8] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. 2012. Touch me once and I know it's you! implicit authentication based on touch screen patterns. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 987–996.
[9] M. O Derawi, P. Bours, and K. Holien. 2010. Improved cycle detection for accelerometer based gait authentication. In *IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. 312–317.
[10] B Dorizzi, R Cappelli, M Ferrara, D Maio, D Maltoni, N Houmani, S Garcia-Salicetti, and A Mayoue. 2009. Fingerprint and On-Line Signature Verification Competitions at ICB 2009. In *International Conference on Biometrics (ICB)*.
[11] A. L. Fantana, S. Ramachandran, C. H. Schunck, and M. Talamo. 2015. Movement based biometric authentication with smartphones. In *IEEE International Carnahan Conference on Security Technology (ICCST)*. 235–239.
[12] L. J. Fennelly. 2003. *Effective Physical Security*. Elsevier Science.
[13] I. Griswold-Steiner, R. Matovu, and A. Serwadda. 2017. Handwriting watcher: A mechanism for smartwatch-driven handwriting authentication. In *IEEE International Joint Conference on Biometrics (IJCB)*. 216–224.
[14] I. Griswold-Steiner, R. Matovu, and A. Serwadda. 2019. Wearables-Driven Freeform Handwriting Authentication. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)* 1, 3 (2019), 152–164.
[15] K. Guk, G. Han, J. Lim, K. Jeong, T. Kang, E. Lim, and J. Jung. 2019. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials* 9, 6 (2019), 813.
[16] C. Huang, Z. Yang, H. Chen, and Q. Zhang. 2017. Signing in the Air w/o Constraints: Robust Gesture-based Authentication for Wrist Wearables. In *IEEE Global Communications Conference (GLOBECOM)*. 1–6.
[17] D. Impedovo and G. Pirlo. 2018. Automatic signature verification in the mobile cloud scenario: survey and way ahead. *IEEE Transactions on Emerging Topics in Computing* (2018).
[18] Diederik P. Kingma and Jimmy B. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).
[19] I. Kononenko. 1994. Estimating attributes: analysis and extensions of RELIEF. In *European Conference on Machine Learning*. Springer, 171–182.
[20] A. Levy, B. Nassi, Y. Elovici, and E. Shmueli. 2018. Handwritten signature verification using wrist-worn devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–26.
[21] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser. 2016. Whose move is it anyway? Authenticating smart wearable devices using unique head movement patterns. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–9.
[22] G. Liang, X. Xu, and J. Yu. 2017. User-authentication on wearable devices based on punch gesture biometrics. In *ITM Web Conf.*, Vol. 11. EDP Sciences, 01003.
[23] C. X. Lu, B. Du, H. Wen, S. Wang, A. Markham, I. Martinovic, Y. Shen, and N. Trigoni. 2018. Snoopy: Sniffing your smartwatch passwords via deep sequence learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–29.
[24] D. Lu, D. Huang, Y. Deng, and A. Alshamrani. 2018. Multifactor user authentication with in-air-handwriting and hand geometry. In *IEEE International Conference on Biometrics (ICB)*. 255–262.
[25] D. Lu, K. Xu, and D. Huang. 2017. A data driven in-air-handwriting biometric authentication system. In *IEEE International Joint Conference on Biometrics (IJCB)*.
[26] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S-M Makela, and HA Ailisto. 2005. Identifying users of portable devices from gait pattern with accelerometers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2.
[27] C. Nickel, T. Wirtl, and C. Busch. 2012. Authentication of smartphone users based on the way they walk using k-nn algorithm. In *IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 16–20.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[29] G. Peng, G.and Zhou, D. T. Nguyen, X. Qi, Q. Yang, and S. Wang. 2016. Continuous authentication with touch behavioral biometrics and voice on wearable glasses. *IEEE transactions on human-machine systems* 47, 3 (2016), 404–416.
[30] A. Primo, V. Phoha, R. Kumar, and A. Serwadda. 2014. Context-aware active authentication using smartphone accelerometer measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 98–105.
[31] EH Rothauser. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics* 17 (1969), 225–246.
[32] D. Schürmann, A. Brüsch, S. Sigg, and L. Wolf. 2017. BANDANA—Body area network device-to-device authentication using natural gAit. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 190–196.
[33] R. Shilkrot, J. Huber, J. Steimle, S. Nanayakkara, and P. Maes. 2015. Digital digits: A comprehensive survey of finger augmentation devices. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–29.
[34] P. Shinde, S. Shetty, and M. Mehra. 2016. Survey of Keystroke Dynamics as a Biometric for Static Authentication. *International Journal of Computer Science and Information Security* 14, 4 (2016), 203.
[35] M. Shoaib, S. Bosch, O. Durmaz Incel, H. Scholten, and P. JM Havinga. 2016. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors* 16, 4 (2016), 426.
[36] Y. Song, Z. Cai, and Z. Zhang. 2017. Multi-touch authentication using hand geometry and behavioral information. In *IEEE Symposium on Security and Privacy (S&P)*. 357–372.
[37] T. T. Um, F. MJ. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *ACM International Conference on Multimodal Interaction*. 216–220.
[38] J. Wang, Y. Hsu, and J. Liu. 2009. An inertial-measurement-unit-based pen with a trajectory reconstruction algorithm and its applications. *IEEE Transactions on Industrial Electronics* 57, 10 (2009), 3508–3521.
[39] J. Yang, Y. Li, and M. Xie. 2015. MotionAuth: Motion-based authentication for wrist worn smart devices. In *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 550–555.
[40] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli. 2019. Security and accuracy of fingerprint-based biometrics: A review. *Symmetry* 11, 2 (2019), 141.
[41] Q. Zhang, H. Li, Z. Sun, and T. Tan. 2018. Deep feature fusion for iris and periocular biometrics on mobile devices. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2897–2912.