Under Pressure: Effectiveness and Usability of the Apple Pencil as a Biometric Authentication Tool

Elina van Kempen
UC Irvine
evankemp@uci.edu

Zane Karl
UC Irvine
zkarl@uci.edu

Richard DeAmicis
UC Irvine
rdeamici@uci.edu

Qi Alfred Chen
UC Irvine
alfchen@uci.edu

***Abstract*—Biometric authentication systems, such as fingerprint scanning or facial recognition, are now commonplace and available on the majority of new smartphones and laptops. With the development of tablet-digital pen systems, the deployment of handwriting authentication is to be considered.**

In this paper, we evaluate the viability of using the dynamic properties of handwriting, provided by the Apple Pencil, to distinguish and authenticate individuals. Following the data collection phase involving 30 participants, we examined the accuracy of time-series classification models on different inputs and on text-independent against text-dependent authentication, and we analyzed the effect of handwriting forgery. Additionally, participants completed a user survey to gather insight on the public reception of handwriting authentication. While classification models proved to have high accuracy, above 99% in many cases, and participants had a globally positive view of handwriting authentication, the models were not always robust against forgeries, with up to 21.3% forgery success rate. Overall, participants were positive about using handwriting authentication but showed some concern regarding its privacy and security impacts.

I. Introduction

The three ways of authenticating are commonly designated as “what you know”, “what you have”, and “what you are”. Examples of these authentication methods are passwords, tokens, and fingerprints, respectively. Each method has specific drawbacks, e.g. passwords can be forgotten, and tokens can be lost.

Biometric-based techniques have been developed as an attempt to address some of these challenges. Biometrics leverage unique physical or behavioral characteristics of individuals to provide a convenient method of authentication. Biometrics are unique to each individual, difficult to steal, and do not rely on the user’s memory. We distinguish static biometrics, usually physical characteristics, and dynamic biometrics, the behavioral characteristics of an individual. Common static biometrics include fingerprint scanning and facial recognition, while dynamic biometrics include voice patterns and facial movements.

Handwriting is a type of biometrics that includes both static and dynamic properties. The visual, written form of a handwriting sample is a static factor, and the way in which someone writes, such as the angle they hold the writing instrument and how hard they press the instrument into the writing surface, is considered a dynamic factor.

Static properties of handwriting have been used consistently, usually in the form of signatures. Signatures are used to verify users on documents such as legal contracts and financial records. However, studies cast doubt on the viability of static handwriting properties alone to stand up to forgery i n the face of modern artificial intelligence technologies [4], [7].

Collecting dynamic data from a handwritten sample requires special hardware, such as a pressure sensor or an accelerometer. Digital pens, such as the Apple Pencil, contain several sensors and provide the sensor data to the developer. These devices can thus be used to provide handwriting authentication. Furthermore, since tablets and digital pens are commercially available, a handwriting authentication system could be very easily implemented, without requiring the design of specific and additional hardware.

Possible applications of live handwriting authentication, using dynamic features, comprise credit card payment signatures, package reception, access control, and electronic signature.

Contribution

In this paper, we investigate whether the Apple Pencil can be an effective biometric authentication tool, using dynamic handwriting authentication. We consider the following four research questions as the focus of our work:

RQ1: How effective is handwriting authentication when using data from the Apple Pencil’s sensors?
RQ2: Which type of input would give the best accuracy when implementing handwriting authentication?
RQ3: What would the enrollment process consist of?
RQ4: How would handwriting authentication be received by users?

For the purpose of data collection, we developed an iOS application. Through this application, we gathered a diverse range of user writing samples, including the user’s name and simple drawings, to compare the effectiveness of these different inputs. After the data collection phase, we examined the accuracy of a time-series classification model on the handwriting authentication task. Finally, we conducted a user survey to gauge the opinions users had of the hardware, handwriting process, and security and practicality of handwriting authentication.

Symposium on Usable Security and Privacy (USEC) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-5-2
https://dx.doi.org/10.14722/usec.2024.23056
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

Through the use of the Apple provided API PencilKit, we were able to directly capture five dynamic handwriting features: force recorded by the pressure sensor in the device, azimuth, altitude, and x and y position of the digital pen. We used these features to build a classification model and authenticate users.

This work makes the following contributions:

- We show that high authentication accuracy can be obtained using the Apple Pencil provided biometric data, and a time-series classification model.
- Different input types are analyzed to determine which will provide both high accuracy and user satisfaction.
- We demonstrate the need to benchmark against forgery, both for security and privacy purposes and to obtain the user's trust in the system.
- User satisfaction with regards to the use of the Apple Pencil, and to the handwriting authentication system, is reported.

Organization

The paper is organized as follows. Related work is presented in Section II, followed in Section III by some background on the Apple Pencil device and on the classification method used. Section IV describes the methodology for data collection, forgeries, and user authentication. Section V and Section VII present evaluation and discussion of our work, respectively. We conclude this paper in Section VIII.

II. Related work

Using one's handwriting as an authentication tool has been thoroughly studied, and a signature is a well-established and accepted way of authenticating a person. Many studies, either in the case of improving an authentication mechanism or in the case of forensic analysis, focus on the static analysis of a handwritten expression, i.e. analyzing the output "image" of the written content. This includes the analysis of the size and shape of the written characters, or the spacing between characters. Dynamic analysis of handwriting requires specific hardware, such as sensors and recording devices. Some features considered by dynamic analysis are the angle of the writing utensil, the amount of pressure applied to the writing surface, the speed at which the user writes, or the number of individual strokes used to create the written sample. Our work targets the dynamic analysis of handwriting on digital devices, so we will not discuss purely static analysis works.

Before the development of tablets and corresponding digital pens, custom-made digital pens were built by embedding specialized hardware such as pressure sensors and lasers into the actual pen [12], [17]. While the systems performed well, users in [12] reported discomfort in handling and manipulating the modified utensil, casting doubt that their resulting handwriting was fully representative of their standard handwriting.

Two signature datasets [21], [23] both contain data from people writing only or mainly using WACOM brand tablets and digital pens. A smaller amount of the data found in these datasets was collected using Samsung tablets with a stylus. [13], [19], [24] evaluated their authentication models on one of the two datasets. A few studies, [5], [18], also using WACOM tablets, collected their own sample data for analysis.

Wijewickrama et al. [22] and Tian et al. [20] use different devices to obtain dynamic handwriting biometric data, respectively smart watches and a Microsoft Kinect. Wijewickrama et al. record wrist movements as the user is writing, and Tian et al. record 3D data points created by participants when signing in the air.

Table I summarizes and compares devices and techniques employed by existing work. Existing work focuses on getting a high accuracy for handwriting authentication: the availability and feasibility of deploying a handwriting authentication framework on the target devices was not reported. To the best of our knowledge, previous work also fails to address the usability of the devised systems, and does not report participants' experiences.

III. Background

A. Apple Pencil

Fig. 1: iPad Air 4th generation and Apple Pencil 2nd generation

The Apple Pencil is a digital pen designed to work with Apple iPads, both pictured in Figure 1. Available since 2015, the Apple Pencil 1 is listed at $99 while the Apple Pencil 2, available since 2018, costs $129. In this work, we used the Apple Pencil 2 only. It weighs 20.7 g and has length and diameter of 166 mm and 8.9 mm, respectively [1]. The Apple Pencil 2 charges by magnetically attaching it to a compatible iPad. In the second quarter of 2023, until April 1st 2023, Apple reported 6.67 million dollars in iPad sales [3]. Apple considers the Apple Pencil as an "accessory", so individual sales of the Apple Pencil are not recorded.

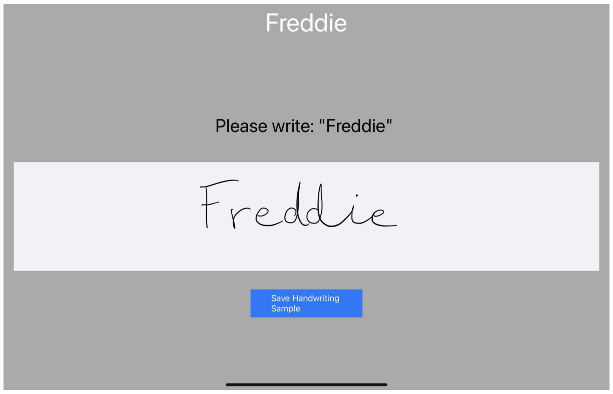
The Apple Pencil 2 provides pressure and tilt sensitivity to the user [1]. Through the UIKit framework, developers can access azimuth, altitude and force data from the Apple Pencil, sent with a 60-240Hz frequency. UIKit also provides the x and y position of the tip of the Apple Pencil on the screen [2].

B. MiniRocket

MiniRocket, [8], is the successor of Rocket, and is one of the fastest and most accurate time series classifier. Rocket's

TABLE I: Comparison of existing work on dynamic handwriting authentication.

	Device	**Input type**	**# Participants**	**Recorded data**
[17]	Homemade digital pen	Digits	12	Acceleration
[12]	Homemade digital pen	Signatures	40	Pressure, inclination, stroke coordinates
[18]	Wacom Art Pad 2 pro Serial	Signatures	14	Pressure, inclination, stroke coordinates
[5]	Wacom tablet	Words	25	Pressure, inclination, stroke coordinates
[20]	Microsoft Kinect	Signatures	18	3D data points
[22]	Sony Smartwatch 3 or LG Watch Urbane	Lowercase and uppercase letters, words	21	Wrist motion
[11]	Apple Pencil	Common password	30	Stroke coordinates
This work	**Apple Pencil 2**	**Names, digits, lowercase and uppercase words, sentences, drawings**	**30**	**Pressure, inclination, stroke coordinates**
Datasets				
[21]	Wacom tablets, Samsung tablets, Samsung phone	Signatures	1526	Pressure, inclination, stroke coordinates
[23]	Wacom Intuos tablet	Signatures	100	Pressure, inclination, stroke coordinates


Freddie

Please write: "Freddie"

Freddie

Fig. 2: A screenshot of the interface of the iPad application for data collection. Here, the user is prompted to write his name.

main contribution was an improvement in computational complexity by transforming input time series data using random convolutional kernels and then using that output to train a linear classifier. We decided to use MiniRocket, due to its speed, which makes it suitable for online handwriting authentication, and its accuracy.

IV. Methodology

A. Ethical considerations

The author's Institutional Review Board (IRB) determined that this study was exempt, under category 3ib. First names were collected from participants for the purpose of analyzing input types, and to avoid the recording of users' signatures. Any other personally identifiable information such as full names, phone numbers, or email addresses, was not collected.

B. Data collection

We used 3 Apple iPad Air devices and the corresponding Apple Pencil 2 to collect data. To collect the data efficiently, we developed an application compatible with iPadOs 16.3.1, which communicates directly with a Google Sheets spreadsheet. The application prompts the user to write or draw some input within a frame, and sends the information to the spreadsheet. Figure 2 shows a screenshot of the application interface.

The study involved a total of 30 participants, all of whom were undergraduate or graduate college students. Participants were given an iPad and an Apple Pencil, and were welcome to hold the iPad as desired when writing, e.g. in their hands, flat on a table, or angled on a table using a support. This decision was made to be able to simulate realistic handwriting authentication using tablets, which are portable and provide the ability to adjust their positioning.

Participants were asked to provide several writing samples using an iPad and an Apple Pencil, and to complete a short user form. In total, these two tasks lasted approximately 15 to 30 minutes to complete, depending on the writing speed of the individual. Participants could take breaks at any time during the procedure. We label each participant with a unique number between 0 and 29. Providing handwriting samples for forgery purposes was voluntary. Six participants agreed and provided samples of their written inputs to be forged.

Each participant provided 21 unique inputs. The input values included both text and drawings. Examples include the participants' first names, the 10 digits, short words like "vegetarian" and "handwriting", and short phrases like "hello world". For all single-word and short-phrase inputs, participants first wrote using all lowercase letters, and then using all uppercase letters. We asked participants to provide their first name instead of a signature for privacy purposes, since participants may not be comfortable with having their actual signature recorded, and we expect participants to be as familiar with writing their first name as they are with writing their signatures. In previous work, researchers ask participants to invent a fake signature and practice until they are comfortable with it [23]. We also had each participant write the phrase "the quick brown fox jumps over the lazy dog", because this sentence includes every character in the English language, and is long. Finally, we asked participants to copy and draw images of a cat, a fish, and a bird.

Each input was repeated five times, totaling 105 samples per participant. This led to an average of roughly 50,000 unique data points recorded from the Apple Pencil per participant. The number of data points had a fairly large range. The maximum number of data points collected from a single participant was over 87,000, while the minimum number was under 27,000. We attribute the large discrepancy to the writing speed and the letter size at which participants recorded their writing samples. With 30 individual users included in our dataset, we obtained a total of 3,150 writing samples.

After data collection, we asked users about their experience, satisfaction, and concerns, in a user survey. The user survey was conducted online via a Google Form, immediately after the participant completed recording all samples.

C. Forgery

We consider the threat model of skilled forgeries as defined by previous work [6], [23]: the attacker has access to the handwritten content needed by the application, previously produced by the victim. That is, the attacker is not an "expert" forger, but can practice as much as needed before trying to authenticate as the victim. "Skilled forgeries" are named in opposition to "random forgeries", where the attacker randomly guesses how the victim is likely to write. Forgers were participants of the study that also provided their handwriting samples.

We provided each forger with handwritten samples from the victim for each of the selected inputs. Forgers were allowed to practice forging the handwriting, for as long as they wanted to. Then, each provided two forged samples per victim handwriting input, i.e. 42 forgeries per victim.

In total, we obtained 10 sets of forgeries on 6 different victims, provided by 5 forgers. We label victims by their participant label, and forgers by a letter. Table II summarizes the acquired forgeries.

TABLE II: Sets of forgeries provided by 5 different forgers, with 6 total victims.

Victim	Forgeries by
0	A, B, C
1	D
2	C, D
5	D
7	C, E
25	D

D. Classification

The data provided by the Apple Pencil and the PencilKit framework consists of a series of points that make up multiple strokes. The iPad receives data from the Apple Pencil every 0.017 to 0.0042 seconds approximately [2]. Figure 3 displays the time series data collected for two different participants writing digits from 0 to 9. Figure 3a displays the altitude, Figure 3b the azimuth, and Figure 3c the force captured by the pressure sensor, for each data point in one sample. In light and dark blue, data collected from participant 0 is shown, in yellow and red, data collected from participant 3 is displayed. Between two samples written by the same individuals, similarities can easily be observed, which illustrate the unique dynamic handwriting characteristics of each person.

As shown in Figure 3, each sample has a different length. Since MiniRocket only works on data of equal length, we first padded the data. We used only the features directly provided by the Apple Pencil, i.e. force, azimuth, altitude, and x and y position of the pen. We decided to only use these features to establish a benchmark on the accuracy of the authentication. Feature engineering may result in better accuracy. The time needed to complete the writing of one sample can also be retrieved, but we chose not to include this measure since it relates directly to the length of the samples, i.e. to the number of data points.

Using sktime [15] and scikit-learn [16], we wrote several Python scripts to perform time-series data classification. Specifically, we used MiniRocket [8] and RidgeClassifierCV.

We built several models, to analyze the effect of input on the accuracy. For instance, we compared a model trained and tested on lowercase text, to a model trained and tested on uppercase text. We also evaluated text-dependent and text-independent authentication, and multiclass and binary models. For each text-dependent authentication model and each participant, 3 out of the 5 samples for each input were selected as part of the training data. The remaining 2 samples were included in the testing data. In the case of text-independent classification, all samples of selected training inputs were included in the training data. The same was done for the testing data.

V. Evaluation

A. Classification accuracy

1) Input comparison: For each input and participant, a classification model decides the label of a test sample within 2 different labels, i.e. decides if the specified participant wrote the sample or not. The results of the 30 models are averaged per input. To train the models, we selected three samples per participant and input, while two samples were reserved for testing purposes. We computed the accuracy and equal error rate (EER) for each model. The EER is averaged across users for each model.

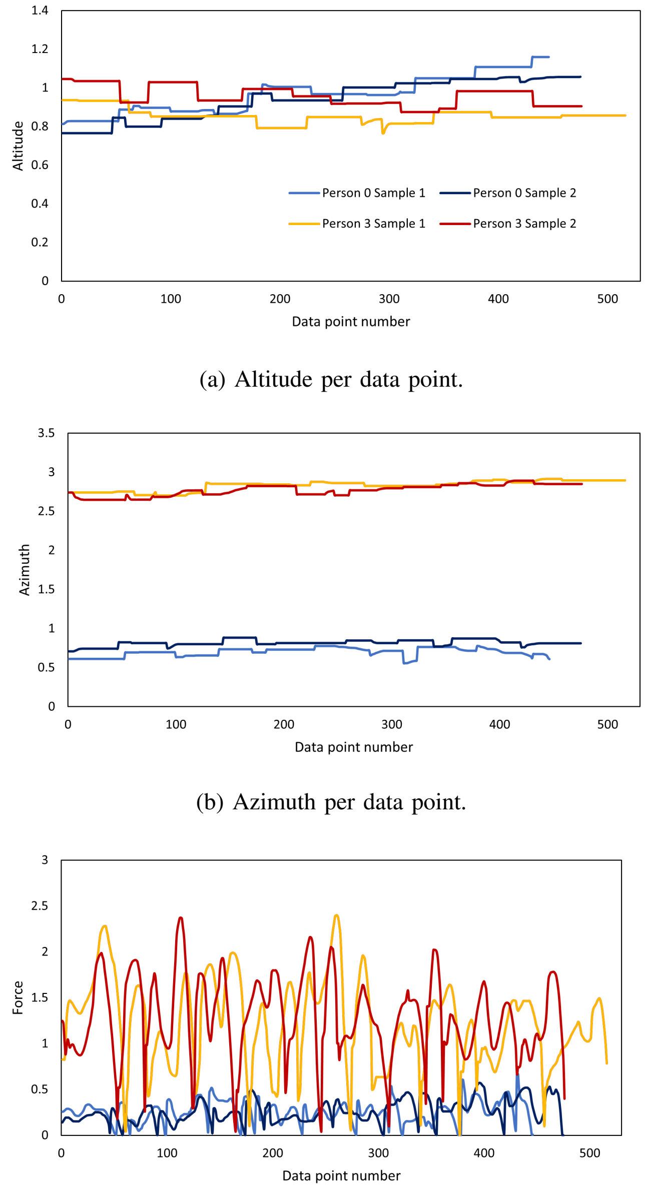
Table III summarizes the accuracy and equal error rate for each different input.

TABLE III: Average accuracy, EER and F1 score for each input

Input #	Input	Accuracy (%)	EER (%)	F1 score
0 -	name	99.9	0.00	0.98
1 -	digits	99.8	0.11	0.97
2 -	carnivorous	99.2	0.52	0.88
3 -	CARNIVOROUS	100.0	0.00	1.00
4 -	vegetarian	99.8	0.06	0.97
5 -	VEGETARIAN	99.6	0.11	0.92
6 -	pineapple	99.7	0.06	0.93
7 -	PINEAPPLE	99.8	0.00	0.97
8 -	handwriting	99.7	0.11	0.93
9 -	HANDWRITING	99.6	0.00	0.89
10 -	security	99.7	0.00	0.93
11 -	SECURITY	99.6	0.17	0.94
12 -	computer	100.0	0.00	1.00
13 -	COMPUTER	99.7	0.29	0.96
14 -	hello world	99.7	0.00	0.94
15 -	HELLO WORLD	99.8	0.06	0.97
16 -	a short sentence	100.0	0.00	1.00
17 -	The quick brown fox [...]	99.9	0.00	0.98
18 -	cat drawing	99.2	0.23	0.83
19 -	bird drawing	99.0	0.06	0.75
20 -	fish drawing	99.4	0.06	0.85

All the models had an accuracy above 99.0%, with the worst performing model, the classification using the bird drawing, had an EER of 0.06% and F1 score of 0.75. Perfect accuracy and F1 score and null EER were observed in 2 of the inputs, namely: "CARNIVOROUS" and "a short sentence". The EER was overall low, and the F1 score was above 0.9 for 16 of the 21 written inputs.

The input consisting of the participant's first name had high accuracy and F1 score, which may be because it was



(a) Altitude per data point.

(b) Azimuth per data point.

(c) Force per data point.

Fig. 3: Data captured by the Apple Pencil for two different users writing digits 0-9, with two samples each.

(almost) unique to each user: most users do not have the same first name. Each user wrote one unique word, instead of all users writing the same word. This is equivalent to the case of handwriting authentication on signatures, where each can be expected to be unique.

The perfect and almost perfect accuracy of the model when using “a short sentence” and the “quick brown fox” sentence as input may be due to the length of the sentences, meaning more data points available to train on.

Using drawings as an input was the least accurate, with models using the drawing of a cat, bird and fish having the lowest accuracy and F1 score, down to a 0.75 F1 score in the case of the bird drawing.

Overall, these results show that a handwriting authentication system that requires each user to write a word 3 times only during the enrollment process would be able to achieve high accuracy. Providing 3 samples takes less than a minute for words and short sentences, for the average person, and on average less than 2 minutes for long sentences and drawings (see Figure 6).

It seems that using words or text is a better option than using simple drawings as an input, to obtain an accurate model.

2) Lowercase vs. uppercase: After splitting lowercase and uppercase words data, we measure the accuracy of models classifying on lowercase text, compared to ones classifying on uppercase text. The selected words are “carnivorous”, “vegetarian”, “pineapple”, “handwriting”, “security”, “computer” and “hello world”.

Models were trained on all words; for each word and each person, 3 samples were picked for training data and 2 for testing.

Table IV displays the accuracy, EER and F1 score for both lowercase and uppercase models. The lowercase-trained classification model had an accuracy of 99.7%, an EER of 0.11% and an F1 score of 0.95, and the uppercase-trained model had an accuracy of 99.6%, an EER of 0.12%, and an F1 score of 0.94.

Both models have similar performance, with the lowercase-trained model having an accuracy and F1 score only 0.1 points higher than the uppercase-trained model.

Comparing input by input, as in Figure III, lowercase-trained models had on average an accuracy of 99.7%, an EER of 0.11% and F1 score of 0.95, while uppercase-trained models had an average accuracy of 99.7%, an EER of 0.10% and F1 score of 0.95. Here again, the models perform similarly.

TABLE IV: Average accuracy, EER and F1 score for lowercase and uppercase input classification

	Accuracy (%)	**EER (%)**	**F1 score**
Lowercase	99.7	0.11	0.95
Uppercase	99.6	0.12	0.94

3) Multiclass vs. binary models: A multiclass classification model is a model that predicts between the 30 classification labels, i.e. one label per participant. A binary classification model is associated with one participant i and outputs 0 or 1: it should predict 1 if the sample was written by i, and 0 otherwise. Thus, one multiclass model is sufficient for classification, and 30 binary models are needed to assess model performance for all participants.

First with a multiclass model, we included all word samples in our evaluation, both lower and uppercase. We did not include sentences, names, digits, or drawings. The accuracy of the model was 99.3% and the EER 0.02%, averaged over 5 training/testing phases.

With binary models, the average accuracy was 99.5% and the EER 0.01%, averaged over all 30 models and over 2 training/testing phases. Figure 4 displays the distribution of the accuracy of individual models. We observe that 20 out of the 30 models have an accuracy above 99.5%.

Binary models demonstrated on average higher accuracy and lower EER than a multiclass classification model.

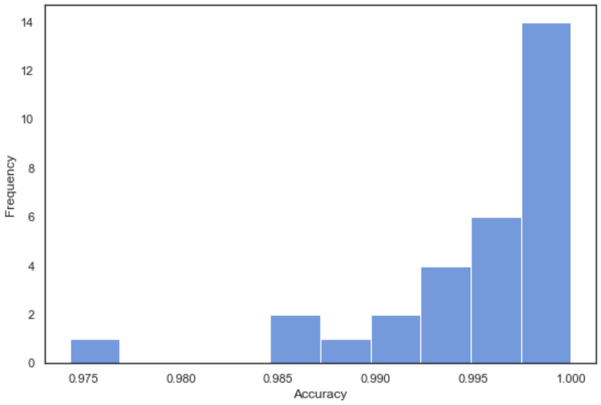


Fig. 4: Accuracy of 30 binary models.

The amount of storage needed for the binary models is similar to the one needed by the multiclass model: the multiclass model has a size of 2,400,302 bytes, and each binary model has a size of 80,882 bytes, for a total of 2,426,460 bytes for all 30 binary models.

However, 30 binary models took longer to train than one multiclass model: approximately 444s for 30 binary models, and 15.3 seconds for one multiclass classification model on a machine with an AMD Ryzen 7 4700U CPU @ 2GHz with 16GB RAM.

4) Text-dependent vs. text-independent: Text-dependent classification indicates that the same text is used for both training and testing of the model, while text-independent classification seeks to predict on text that is different than the text available for training. Text-independent classification may be useful in some applications, e.g. continuous authentication.

In this experiment, we allowed models to train on 4 lowercase words, “carnivorous”, “vegetarian”, “pineapple”, and “handwriting”, and tested on 3: “security”, “computer”, and “hello world”. We repeated this analysis on the same words, in uppercase.

Table V displays the accuracy, EER and the F1 score for each model. Models had a high average accuracy, with an average accuracy of 98.4% and 98.3% for lowercase and uppercase, respectively, and both EERs of 0.08%. However, the F1 scores were lower, with values of 0.63 and 0.61, showing that text-independent classification performs worse.

In general, text-dependent classification models, presented in previous sections of this paper, were more accurate than text-independent classification models. For instance, in Section V-A3, both binary and multiclass classification models have an accuracy above 99.3%.

B. Forgery

We studied the effect of skilled forgery on classification. A skilled forgery is defined in opposition to random forgery: a skilled forger has access to the victim’s handwriting samples and can practice, while a random forger just guesses a person’s handwriting.

TABLE V: Average accuracy, EER and F1 score for text-independent models.

	Accuracy (%)	EER	F1 score
Lowercase	98.4	0.08	0.63
Uppercase	98.3	0.08	0.61

TABLE VI: Fraction of successful forgeries per victim.

	Fraction of successful forgeries
Forgery of 0	21.3%
Forgery of 1	2.4%
Forgery of 2	7.4%
Forgery of 5	0.0%
Forgery of 7	0.0%
Forgery of 25	0.0%

We obtained 10 sets of forgeries, for 6 participants. In total, 5 different forgers provided the 10 sets of forgeries. Each forger had 2 tries per input, resulting in 42 forged data samples per victim.

Forgeries were tested using the models trained on each input for every victim: for each input, we trained binary models including all the data collected from all participants except the forgers, and then predicted on the forged data. We consider a forgery “successful” if the predicted label corresponds to the victim’s label.

Every forgery attempt was tested 5 times, the average rate of success is reported. Table VI presents the average percentage of successful forgeries for each writing input, out of all data samples for each victim.

The success of each forgery was very dependent on the victim. For example, forgeries of participants 0, 1, and 2 had a high success rate, with a minimum success rate of 2.4% and a maximum of 21.3%, while no forgery attempt was successful on participants 5, 7 and 25. This may mean that some users’ handwriting is easier to forge, some forgers are more skilled than others, or that the model did not perform well for a specific user.

It is interesting to note that forgers needed more time to complete each writing sample than their victims. Figure 5 shows the time needed to complete 42 handwriting samples, 2 per input, for each victim, and for each forger for the targeted victim.

Since forgers are slower than their victims, implementing a timed defense may be an option. However, it should not prevent honest users from authenticating, if they happen to be slower than usual at a certain time.

C. User study

In this section, we report participants’ answers to the user survey, which they completed after data collection.

1) Satisfaction with the Apple Pencil: Initial questions aimed to assess the participants’ opinions on the Apple Pencil

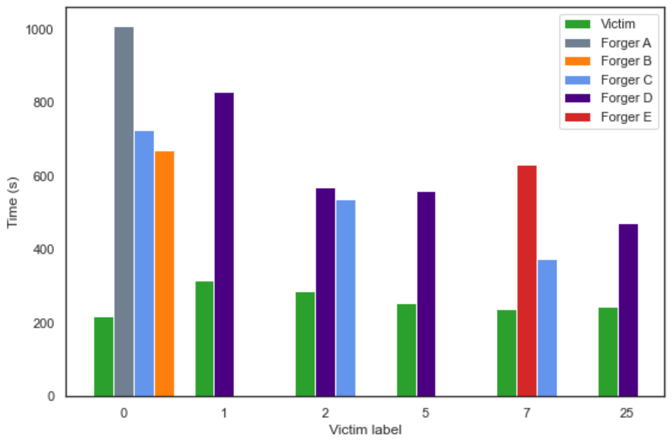


Fig. 5: Time to complete writing 42 samples, for each victim.

device. First, participants shared their satisfaction with using the Apple Pencil, rating the following statements:

- "The Apple Pencil was easy to use."
- "I experienced no physical discomfort or strain when using the Apple Pencil."

Participants' answers are shown in Figure 7. The majority of users considered the Apple Pencil easy to use, with only 2 of the 30 participants that did not find the device simple to control. One participant commented that the digital pen "flowed really well, felt like a regular pencil", another "could rest my hand on the glass just like a real pencil". One participant did comment "it's slippery". Note that 9 of the 30 participants claim to never use an Apple Pencil, and 6 had only tried it a few times. Out of these 15 participants, only 1 found the Apple Pencil hard to handle.

Even though the digital pen was reported easy to use, 4 participants experienced some physical discomfort or strain when using the Apple Pencil, and 7 did not find that using the Apple Pencil was particularly comfortable. This may be due to the prolonged and constant use of the pen during data collection, as well as the tiredness due to writing the same text multiple times. Specifically, users commented that "it is heavier than the usual pen", "slightly more tiring than physical paper and pen", and "hand started to cramp".

On average, participants who did not find the pencil comfortable applied more force on the Apple Pencil, with an average force of 1.15, for participants who disagreed that the use of the pen was comfortable throughout, and one of 0.84 for those who strongly agreed.

2) User-friendliness of writing authentication: When presented with the statement that "handwriting authentication is easy to use", the majority of the participants, 21/30, agreed or strongly agreed. Participants' comments include: "Handwriting is easy so is authentication" and "Felt as easy as any other system". One participant disagreed with the statement, and commented: "Not that convenience".

Similarly, most of the participants, again 21/30, thought that writing a few words of sentences was fast. Two participants disagreed, while three had a neutral point of view.

We then asked participants which input type was their favorite, in a situation where they would use handwriting authentication. Figure 8 displays participants' answers, with the majority of participants preferring words or doodles. A participant that enjoyed using words as input better expressed that "doodles were too hard, sentences made my hand cramp", while another with doodles as the preferred input type stated that "they were more fun".

As shown in Figure 6, lowercase words were the fastest to write, which may be why participants preferred writing words. Inputs that took the longest to write or draw were the long sentence and the doodles. We hypothesize that participants enjoyed the doodles for the fun aspect.

3) Usefulness of handwriting authentication: Later questions focused on the perception of the usefulness of handwriting authentication. 21 participants agreed or strongly agreed that they would use handwriting authentication in their daily life, if given the opportunity. A participant wrote that handwriting authentication would be "great for large purchases". On the other hand, one participant commented "No secure. I don't trust it". Five participants stated that they would not use handwriting authentication if it were available.

Participants then rated the usefulness of handwriting authentication for specific applications: electronic document signatures, credit card payment signatures, package reception, essay writing, art, and access control. Figure 9 displays the results, and shows that using handwriting authentication for signatures, for electronic documents and credit card payments, was identified as very or somewhat useful by 28 and 26 out of 30 participants, respectively. Implementing handwriting authentication for essay writing did not seem useful for 11 of the 30 participants.

4) Privacy and security concerns: Finally, we asked participants about any privacy and security concerns they had if handwriting authentication were to become available.

Participants' answers are shown in Figure 10. Overall, participants thought that handwriting authentication is more secure than a traditional signature, but many remained neutral. When asked if they felt that using handwriting biometric data was secure, the majority also chose to remain neutral. This may mean that participants were unsure of the privacy and security properties of handwriting authentication, and how it would or could be used. Similarly, while 13/30 participants were concerned that someone would forge their handwriting, 11 responded neutrally.

Still, only 6 out of the 30 participants were not concerned that someone would forge their handwriting. Since forgery seems to be a common concern, any implementation of handwriting authentication should benchmark robustness against forgery.

Ultimately, when asked if they had any privacy or security concerns regarding the use of handwriting authentication, many participants answered positively. We list below participants' answers to the question "Do you have any privacy or security concerns regarding the use of handwriting authentication technology?":

- "Don't want my identity stolen"

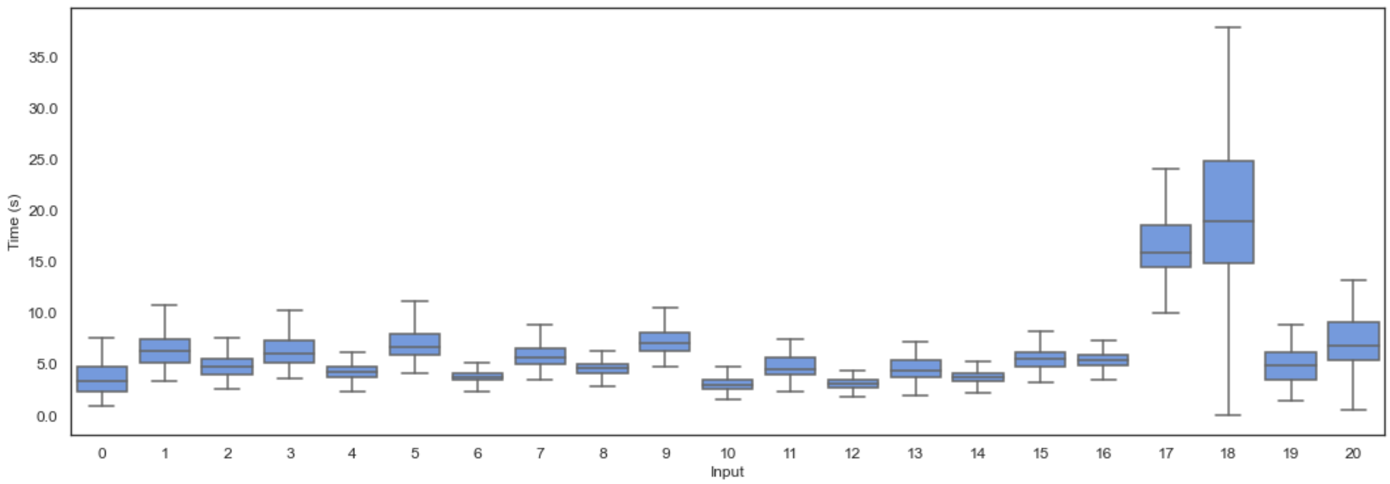
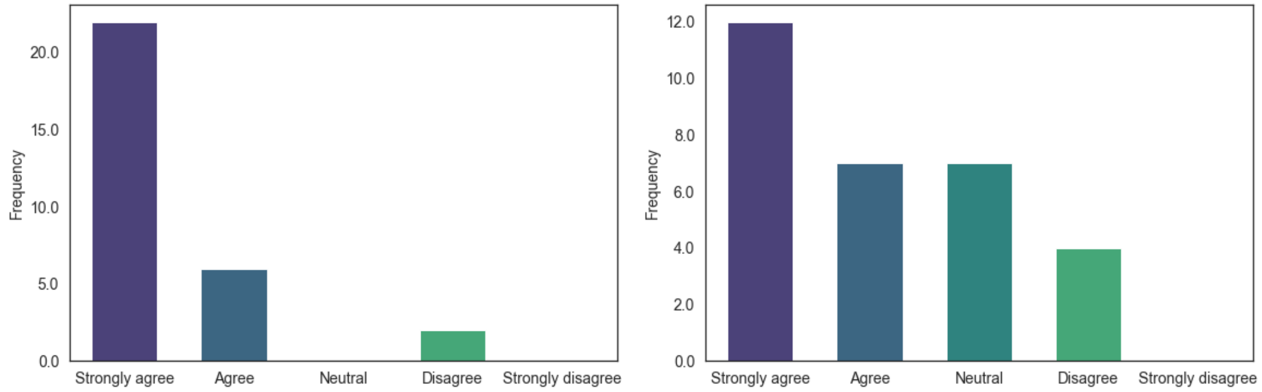


Fig. 6: Time required for participants to write each different input. Inputs are given by the input numbers, detailed in Table III.



(a) “The Apple Pencil was easy to use.”

(b) “Using the Apple Pencil felt comfortable throughout.”

Fig. 7: User satisfaction concerning the use of the Apple Pencil.

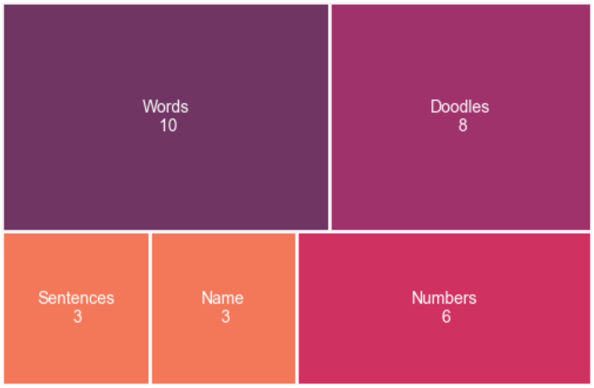


Fig. 8: Answers to: “Which type of input did you prefer when writing?”

- “I would like to have professionals evaluate the security and show the results.”
- “Yes, how easy to forge is it?”
- “I do not have anything specific but I feel like there might be some problems.”
- “Yes, handwriting can be forged”
- “Yes” (twice)
- “Some privacy concerns”
- “Hacking”
- “Ppl’s handwriting change from time to time”
- “Handwriting may change ”
- “My concerns are whether the technology would accurately get my signature from a forger”
- “No, as long as the hypothetical pressure fingerprint is stored securely.”

17 of the 30 participants answered “No” to the question.

VI. Results Summary and Answers to Research Questions

RQ1: *How effective is handwriting authentication when using data from the Apple Pencil’s sensors?:* As shown in Section V, many models had accuracy above 99%, an EER lower than 0.5%, and an F1 score above 0.9, which seems

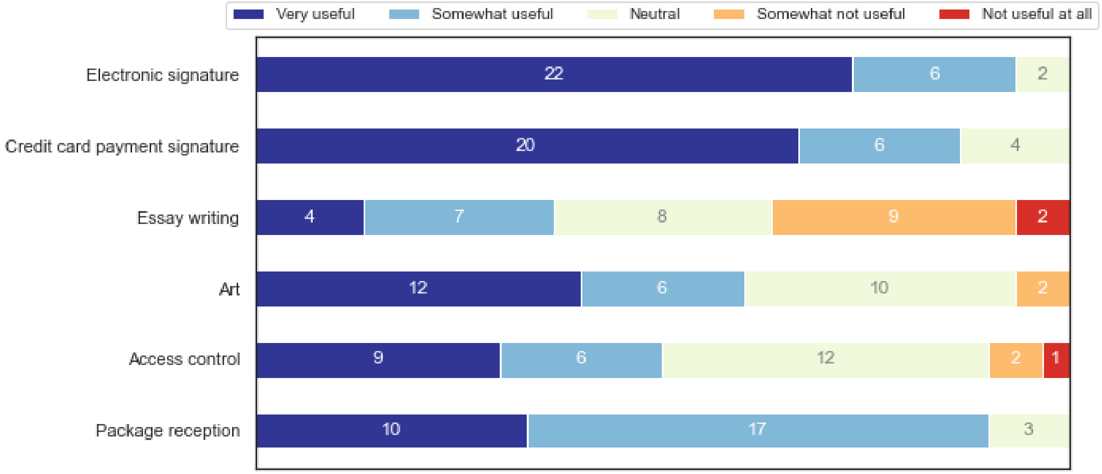
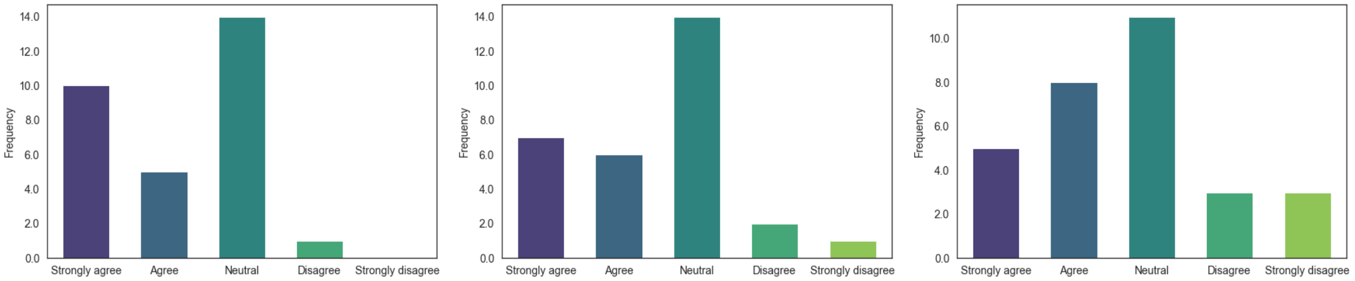


Fig. 9: Answers to: “For the following applications, rate how useful handwriting authentication would be.”



(a) “Compared to a traditional signature, handwriting authentication is more secure.”

(b) “I felt that using handwriting biometric data is very secure.”

(c) “I am concerned that someone would forge my handwriting.”

Fig. 10: User security and privacy concerns.

very promising for a handwriting authentication system. This was achieved using only the data directly provided by the Apple pencil sensors, i.e. force, azimuth, altitude, and x and y position of the pencil. With more data manipulation, feature engineering, and using different classification techniques, a higher accuracy could be achieved. Thus, the Apple Pencil could be used as a device to help implement handwriting authentication.

RQ2: *Which type of input would give the best accuracy when implementing handwriting authentication?:* Simple drawings/doodles seemed to be less accurate than actual handwriting, as shown in Section V-A1, thus we do not recommend them to be used for handwriting authentication at this time. Names/signatures, which are unique to each user, had high accuracy, along with longer sentences. A drawback of using sentences as input is that the user may get tired of writing for longer periods of time, during both enrollment and later authentication. Using either lowercase or uppercase text did not make a significant difference. Most words as input also had a reasonably high accuracy, and could be used to authenticate users. Text-independent classification was less accurate than text-dependent classification, but may have specific applications that require it.

RQ3: *What would the enrollment process consist of?:* Participants found that writing a few words was fast, stated in section V-C2, and as shown in section V-A1, training a model with only 3 training samples per user still results in a model with accurate predictions. We can conclude that such enrollment would represent a minor inconvenience to users, although as with all biometric systems, enrollment does require the user’s physical presence. If more samples per user are wanted for a higher accuracy classification, users may get tired and feel discomfort from writing for a longer amount of time (section V-C1).

RQ4: *How would handwriting authentication be received by users?:* On average, users found handwriting authentication useful, and claimed that they would use it in a real world setting (Section V-C3). Participants also thought it was easy to use (Section V-C2). However, presented in Section V-C4, many were concerned or unsure about the security and privacy guarantees of a digital pen-enabled handwriting authentication system. A thorough evaluation of such a system is needed against forgery and other attacks. In addition, some users were concerned about the safe storage of their handwriting biometric data.

VII. Discussion and limitations

A. Dataset

Several factors may impact our results. First, during the data collection phase, users sometimes verbally reported fatigue and boredom, because of the repetitiveness of the task. Because of this, they may start writing faster or with less precision toward the end of the data collection process. Users who did not use an Apple Pencil or any similar digital pen before may have been slower at the beginning of the data collection. Some users also paused for several seconds while writing a word. The number of participants in our dataset may further influence our results, and obtaining a larger dataset is a future goal.

B. Forgery

The forgery success rate was high for some victims, notably a success rate of 21.3% for the first victim. However, for three participants, forgeries were never successful. This may be attributed to several factors. For instance, some individuals may have a more consistent handwriting that is easier to replicate, and individuals from the same cultural or national background may have been taught a specific handwriting style and have an easier time forging that handwriting style. Some forgers invested more time on their forgeries (Figure 5), which may have provided a better result. Additionally, it is possible that the limited number of data samples available could cause the models to overfit on some participants.

The high forgery success rate is concerning, and real-life forgers that are more motivated, have more experience, and spend more time practicing forgery than the study participants could get an even higher success rate than what is reported in this paper. Implementing defenses is necessary before deploying a handwriting authentication system. Enhancing the system's robustness against forgery could involve setting a time limit or considering more time-related features, such as the duration of a stroke, the interval between strokes, and the overall speed of writing. These measures could help decrease the forgery success rate by forcing adversaries to write faster. The classification models could also be further trained for robustness using reinforcement learning and feedback mechanisms. Using ensemble methods and multiple classifiers may also improve the results, as well as establishing an appropriate classification threshold.

C. Deployability and scalability

We found that 57% of the participants used a digital pen occasionally to regularly, and thus were already familiar with the system. Note that this study took place in a college setting and this number may not be representative of the overall population. Most participants had only ever tried the Apple Pencil, and 4 participants only had used a different brand of digital pen at least once, e.g. HP Pen and Microsoft Surface Pen. 70% of the participants found writing a few words to be fast, indicating that writing authentication would not be an overbearing burden on users. Thus, the users' pre-existing understanding of the hardware and the convenience of writing would ease the deployment of the system.

Besides, the time needed to transform and classify 600 data samples was 7.61 seconds, giving an average of 12.68 milliseconds per data sample, which appears sufficient for real-world use. However, the size of the classification model increases as a function of the number of users, increasing the storage needs for large datasets.

This raises concerns about the scalability of the system. To address this, model compression and quantization can reduce the storage needs. Using a system-owned dataset with a fixed size for training user models or selecting a random subset of users as the training dataset can significantly reduce the time needed to train each model. Implementing periodic re-training of user models can also prevent the need for retraining every time a new user is added.

Lastly, the price of the Apple Pencil, along with the required iPad, can be a factor that would prevent some individuals or organizations from deciding to use handwriting authentication.

D. Privacy challenges

Handwriting authentication may appear more privacy-friendly than other biometric systems: no need to scan a face, an iris, or a fingerprint. The system does not know what the user looks like. However, sensitive and personal information can be inferred from one's handwriting, such as gender, health, or emotional state [9], [10], [14].

In addition, the secure storage and communication of handwriting biometric data must be assured by any entity implementing handwriting authentication.

E. Physical limitations and accessibility

Obvious physical limitations to handwriting authentication are injuries and disabilities, e.g. a person may have a broken wrist or arm and be unable to write for some short or long period of time, or have a permanent disability preventing them from having a distinct handwriting.

Additionally, some people do not know how to write, and will be unable to authenticate in this way.

A person's handwriting may also change with time, requiring regular re-enrollment. This is especially true for children, whose handwriting changes as they acquire more dexterity. A timeline for periodic re-enrollment would need to be estimated.

VIII. Conclusion and future work

We evaluated the performance of handwriting authentication on different inputs and parameters using an Apple Pencil 2, and analyzed participants' view on handwriting authentication using smart pens and digital tablets. Handwriting authentication was most accurate on text inputs, and performed best in a text-dependent scenario, with an accuracy above 99% and F1 score above 0.9 for most cases. Participants were positive about using handwriting authentication, but many were concerned about possibilities of handwriting forgery and the security implications of such system.

While high accuracy is achieved for handwriting authentication, further evaluation of the models' robustness against forgeries is necessary before deploying any handwriting authentication system. Furthermore, both topics of forgery by AI or deep learning models and adversarial attacks on handwriting authentication models need more research.

Future work will aim at expanding our dataset to allow additional contributions. We expect future research in this area to work toward improving the handwriting classifying models by performing feature extraction and selection.

Inspired by handwriting authentication, we intend to explore the possibility of authenticating digital art using the Apple Pencil's properties. An entire artistic project will likely provide distinguishing features corresponding to the given artist. In this way, digital art might be "signed" by virtue of its style and method.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their valuable feedback, and the study participants for their time.

REFERENCES

[1] Apple, "Buy apple pencil 2," https://www.apple.com/shop/product/MU8F2AM/A/apple-pencil-2nd-generation.

[2] AppleDeveloperDocumentation, "Handling input from apple pencil," https://developer.apple.com/documentation/uikit/pencil_interactions/handling_input_from_apple_pencil.

[3] AppleInc., "Condensed consolidated statements of operations," https://www.apple.com/newsroom/pdfs/FY23_Q2_Consolidated_Financial_Statements.pdf, May 2023.

[4] L. Ballard, D. Lopresti, and F. Monrose, "Evaluating the security of handwriting biometrics," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[5] N. Begum, M. A. H. Akash, S. Rahman, J. Shin, M. R. Islam, and M. E. Islam, "User authentication based on handwriting analysis of pen-tablet sensor data using optimal feature selection model," *Future Internet*, vol. 13, no. 9, p. 231, 2021. [Online]. Available: https://doi.org/10.3390/fi13090231

[6] K. Bibi, S. Naz, and A. Rehman, "Biometric signature authentication using machine learning techniques: Current trends, challenges and opportunities," *Multimedia Tools and Applications*, vol. 79, no. 1-2, pp. 289–340, 2020.

[7] J. Chen, "Handwritten biometric systems and their robustness evaluation: a survey," in *Technical Report*. Citeseer, 2012.

[8] A. Dempster, D. F. Schmidt, and G. I. Webb, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B. C. Ooi, and C. Miao, Eds. ACM, 2021, pp. 248–257. [Online]. Available: https://doi.org/10.1145/3447548.3467231

[9] M. Faundez-Zanuy, J. Fierrez, M. A. Ferrer, M. Diaz, R. Tolosana, and R. Plamondon, "Handwriting biometrics: Applications and future trends in e-security and e-health," *Cognitive Computation*, vol. 12, pp. 940–953, 2020.

[10] M. Faundez-Zanuy and E. Sesa-Nogueras, "Preliminary experiments on automatic gender recognition based on online capital letters," in *Recent Advances of Neural Network Models and Applications: Proceedings of the 23rd Workshop of the Italian Neural Networks Society (SIREN), May 23-25, Vietri sul Mare, Salerno, Italy*. Springer, 2014, pp. 363–370.

[11] J. K. Han, B. Kang, and D. Wong, "HWAuth: Handwriting-based socially-inclusive authentication," in *SIGGRAPH Asia 2021 Posters, Tokyo, Japan, December 14-17, 2021*, S. J. Shiota, A. Kimura, and W. A. Ma, Eds. ACM, 2021, pp. 28:1–28:2. [Online]. Available: https://doi.org/10.1145/3476124.3488638

[12] C. Hook, J. Kempf, and G. Scharfenberg, "A novel digitizing pen for the analysis of pen pressure and inclination in handwriting biometrics," in *Biometric Authentication, ECCV 2004 International Workshop, BioAW 2004, Prague, Czech Republic, May 15, 2004, Proceedings*, ser. Lecture Notes in Computer Science, D. Maltoni and A. K. Jain, Eds., vol. 3087. Springer, 2004, pp. 283–294. [Online]. Available: https://doi.org/10.1007/978-3-540-25976-3_26

[13] A. A. Jaini, G. Sulong, and A. Rehman, "Improved dynamic time warping (DTW) approach for online signature verification," *CoRR*, vol. abs/1904.00786, 2019. [Online]. Available: http://arxiv.org/abs/1904.00786

[14] L. Likforman-Sulem, A. Esposito, M. Faundez-Zanuy, S. Clémençon, and G. Cordasco, "Emothaw: A novel database for emotional state recognition from handwriting and drawing," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 2, pp. 273–284, 2017.

[15] M. Löning, A. J. Bagnall, S. Ganesh, V. Kazakov, J. Lines, and F. J. Király, "sktime: A unified interface for machine learning with time series," *CoRR*, vol. abs/1909.07872, 2019. [Online]. Available: http://arxiv.org/abs/1909.07872

[16] 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] R. Renuka, V. Suganya, and B. Arun Kumar, "Online handwritten character recognition using digital pen for static authentication," in *2014 International Conference on Computer Communication and Informatics*, Coimbatore, 2014, pp. 1–5.

[18] D. Sakamoto, M. Kondo, H. Morita, D. Muramatsu, M. Sasaki, and T. Matsumoto, "Dynamic biometric person authentication using pen signature trajectories," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, vol. 4. IEEE, 2002, pp. 2078–2082.

[19] M. Saleem and B. Kovári, "K-nearest neighbour and dynamic time warping for online signature verification," *CoRR*, vol. abs/2111.14438, 2021. [Online]. Available: https://arxiv.org/abs/2111.14438

[20] J. Tian, C. Qu, W. Xu, and S. Wang, "Kinwrite: Handwriting-based authentication using kinect," in *20th Annual Network and Distributed System Security Symposium, NDSS 2013, San Diego, California, USA, February 24-27, 2013*. The Internet Society, 2013. [Online]. Available: https://www.ndss-symposium.org/ndss2013/kinwrite-handwriting-based-authentication-using-kinect

[21] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Deepsign: Deep on-line signature verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 229–239, 2021.

[22] R. Wijewickrama, A. Maiti, and M. Jadliwala, "Write to know: on the feasibility of wrist motion based user-authentication from handwriting," in *WiSec '21: 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks, Abu Dhabi, United Arab Emirates, 28 June - 2 July, 2021*, C. Pöpper, M. Vanhoef, L. Batina, and R. Mayrhofer, Eds. ACM, 2021, pp. 335–346. [Online]. Available: https://doi.org/10.1145/3448300.3468290

[23] D.-Y. Yeung, H.-Y. Chang, Y. Xiong, S. E. George, R. S. Kashi, T. Matsumoto, and G. Rigoll, "SVC2004: First international signature verification competition," in *Lecture Notes in Computer Science*. Springer Science+Business Media, 2004, pp. 16–22.

[24] F. J. Zareen and S. Jabin, "Authentic mobile-biometric signature verification system," *IET Biometrics*, vol. 5, no. 1, pp. 13–19, 2 2016. [Online]. Available: https://doi.org/10.1049/iet-bmt.2015.0017