Wearables-Driven Freeform Handwriting Authentication

Isaac Griswold-Steiner, Richard Matovu, and Abdul Serwadda

***Abstract*—With the ubiquity of handwriting in everyday tasks, it is surprising that existing avenues for handwriting authentication remain largely out of reach for the average individual or organization. Current solutions often rely on expensive or specialized equipment, and most existing research focuses on signatures rather than freeform handwriting. This limits the applicability of such technology to a narrow range of scenarios. In this paper we argue that wearable devices might make handwriting authentication scalable and affordable.**

We design and evaluate two wearables-driven freeform handwriting authentication systems, one centered on a deep neural network and the other using human-engineered features. Our authentication systems are thoroughly tested across three writing experiments (involving 53 participants) that were carefully mapped to typical writing scenarios.

We show the best performing configuration to attain an Equal Error Rate of 5.51%, suggesting the potential of this modality for use in a multi-modal authentication system. To evaluate how our authentication systems perform against attacks by determined attackers, we developed and evaluated two impostor attacks that correspond to highly likely attack vectors. We then show that certain authentication system configurations are resistant to the attack. This paper represents an important step towards consumer ready wearables-driven freeform handwriting authentication.

***Index Terms*—behavioral biometrics, wearables, handwriting, authentication, impersonation attacks**

I. Introduction

As the wearable device market continues to show rapid growth, users will increasingly find these devices integrated with their daily lives. This trend makes wearable devices a strong platform for building applications and tools that improve quality-of-life. These often include fitness and health monitoring applications; but wearable devices also offer a opportunity to leverage daily user behavior for digital user authentication and a wide variety of security enhancements.

Despite the wide range of use-cases in which handwriting offers a convenient means to verify user identify, there is not yet a reliable, fully automated handwriting authentication technology that breaks into the mainstream. The current state-of-the-art methods face a number of challenges, including usability issues, applicability to a limited set of writing scenarios, and high costs for the hardware needed to drive the authentication process.

For example, one of the most researched types of automated handwriting authentication involves the scanning of written text to digitize it before applying image processing and authentication algorithms (e.g., see [1], [2], [3]). While this approach has been shown to demonstrate promise, it has a number of drawbacks, including high costs due to the need for additional hardware (i.e., scanners) and poor usability (i.e., the scanning process is cumbersome and time consuming).

Another commonly fronted scheme uses touch-sensitive tablets or computers which allow authentication technology to leverage the pressure and location of writing. However, this assumes that the user is writing on a tablet and it does not apply to other writing surfaces. This means that if a user wanted to authenticate their handwriting on a different surface (e.g., a piece of paper or whiteboard), they would not be able to use the technology (see Section II for a more detailed description of the state-of-the-art). In this paper, we argue that wearable devices (i.e., watches, fitness trackers and other sensor-enabled wrist-worn devices) might hold the key to a practical, user-friendly, inexpensive, yet reasonably accurate handwriting authentication method. Relative to the state-of-the-art schemes described above, wearable devices offer a number of advantages, including the following:

1) Unobtrusive Experience – Once a user were to setup the handwriting authentication system on a smartwatch, the authentication process should require little to no intervention. It should be less time consuming and distracting than methods which require document scanning every time the authentication process occurs, *2) Multi-Purpose Capabilities and Broad Appeal* – Handwriting authentication is just one of many possible applications for a wrist-worn device (others being fitness, health, weather, or email apps). In contrast with a smartpen, the multitude of potential uses of a smartwatch will make it more likely that a user continues to utilize the device and any applications on it, *3) Environmental Flexibility* – Relying on a smartwatch to gather writing data could facilitate the training of a system for a wide range of writing surfaces and scenarios (e.g., on a desk, chalkboard, whiteboard, or on a book). All that would be required is the collection of data for the particular surface. This is in contrast with a tablet or phone where the writing is limited to a touch screen.

Motivated by the above described benefits of wearables, we designed and evaluated a *wearables-driven freeform handwriting authentication* system. The core mechanism behind our scheme is that the wrist movements executed while a person writes freely provide a movement fingerprint that captures the unique properties of a user's handwriting. Using accelerometers embedded in the watch, this signature can be captured and used to train machine learning classifiers which could then be used to authenticate the user's handwriting. To the best of our knowledge, we are the first to study the use of a wrist-worn

Manuscript received July 9, 2018.

This research was supported by National Science Foundation Award Number: 1527795.

I. Griswold-Steiner, R. Matovu, and A. Serwadda are with the Department of Computer Science at Texas Tech University in Lubbock, TX 79409, USA.

device for the authentication of a user's freeform handwriting (see details of related schemes in Section II). The contributions of this paper are summarized below:

1) **Wearables-Driven Freeform Handwriting Authentication**: We designed and evaluated a freeform handwriting authentication system that uses wrist-worn devices to capture hand movement patterns. To evaluate the potential of this approach we collected data for three experimental conditions: (1) users copied a specific prompt onto a new piece of paper, (2) users copied a unique piece of writing, and (3) users responded to essay-style prompts. Each of these tested different types of writing scenarios that might be encountered in the real world. Our best scenario (E1) has an Equal Error Rate of 5.51%. While this Equal Error Rate (EER) might not be suited for a standalone authentication system, it could be used as part of a multi-modal authentication system.
2) **Systematic Exploration of how Feature Learning Improves on Traditionally Engineered Features for Handwriting Authentication**: To leverage recent advances in feature-learning based algorithms, we developed an authentication system using a deep neural network (DNN). To analyze the relative performance of feature learning against human-engineered features, we also used a set of traditionally engineered features for authenticating users in all experiments. We then explored a large number of configurations for the authentication systems and fusion strategies for boosting their predictive power. These results show us the types of scenarios in which our systems would be most useful based on the results and the amount of data required. The best configurations were then reused during the attack phase, giving us insight into the differences in how each configuration reacted to trained impostor attacks. Our results show that a combination of feature paradigms is likely the best option for creating accurate and secure wearables-driven freeform handwriting authentication.
3) **Performance Evaluation Against Sophisticated Impersonation Attacks**: For two different authentication systems based on learned features and human engineered features we evaluated the impact of two types of impersonation attacks: (1) where an adversary has access to handwritten documents for their target, providing visual information on handwriting characteristics and (2) where the attacker additionally has video of their target writing, allowing them to get insight into the motion dynamics of the victim's writing. We found that an attack using video of the target writing performed better than one who observed handwritten documents of the victim. Further, we show that there are significant differences in the performance of each authentication system against the attacks.
4) **Large Wearable Device Handwriting Dataset with Trained Forgeries**: We present a methodically collected dataset that captures handwriting motion and orientation patterns during freeform writing from 50-53 users (depending on the experiment). The dataset includes three different kinds of writing experiments which represent typical writing scenarios and non-zero-effort forgeries conducted by impersonators who meticulously trained to match their victim's writing pattern. Forgeries map to two likely attack vectors that a freeform handwriting system might face, namely, forgeries based on written documents from the target and videos of their writing process. This dataset has the potential to support a wide range of research streams in areas such as authentication system design, impostor attack understanding, and, the connections between handwriting and cognitive load. As far as we are aware, there exists no such public dataset currently[1].

Road-map: The rest of the paper is organized as follows. We review related work in Section II. Next, we describe our data collection experiments and authentication systems in Sections III and IV, present the authentication results in Section V, and explain our threat model and attack results in Section VI. We finally provide an overall discussion and conclusions for this work in Section VII.

II. Related Work

This paper builds upon our previous conference paper [12], where we showcased an authentication system for freeform handwriting using wearable devices. The current work differs from the previous conference paper in the following significant ways: (1) we use a more stringent threat model, (2) we implement a feature learning scheme and evaluate its performance relative to traditional feature engineering approaches, and, (3) we use a much larger user population that provides a more realistic representation of how the system might perform in practice. See a detailed description of these extensions to the paper in our contributions, discussed in Section I. Besides our own previous study [12], there are two relevant streams of related research: (1) user authentication with writing and (2) spoof attacks on writer authentication systems. We discuss these streams of research next, a summary of these streams of research is also available in Table I.

A. Handwriting Authentication

Research on handwriting authentication can be broadly categorized into motion-sensor oriented (e.g., see [6], [7]) and image processing-based (e.g., see [2], [3], [5]). Due to the significant differences between our work and that of image processing-based techniques (recall discussion in Section I), we do not describe this line of research further. Here we focus our discussion on motion-sensor oriented handwriting authentication, which is additionally divided into that which relates to signatures or freeform handwriting (e.g., written essays). The following subsections are devoted to describing these streams of research further.

[1]The dataset can be downloaded from the following link [4]. We are publishing this dataset in the form of a PostgreSQL database with tables for different types of experiments and the raw data itself.

TABLE I
OVERVIEW OF DIFFERENCES BETWEEN OUR WORK AND RELATED RESEARCH TRAJECTORIES.

Related Research	Example Papers	How Our Work Differs From Related Research
Using images of written text for user authentication.	[2] [3] [5]	Our method uses motion and orientation sensor data collected by wearables. Our method thus does not require the laborious scanning process that all image-based methods involve. Hence our method is user friendly, usable in real-time applications, and scalable to concurrently serving large numbers of users.
Signature authentication using sensor data (e.g., wearable, tablet, etc).	[6] [7] [8]	A signature is a short and highly practiced segment of text which is relatively consistent over time and is typically verified by matching segments of text. Our work targets freeform handwriting patterns and is not tied to specific written content, posing a different pattern recognition problem.
Authentication with freeform handwriting using sensor data (e.g., from tablets, smart pens, etc).	[8] [9]	The majority of papers used tablets for gathering freeform handwriting data, giving detailed spatial, temporal, and kinematic information. The kind of wrist-worn device used in this research provides much less information, as it only captures motion patterns of the wrist, a location that is far away from the writing action.
Trained forgeries on handwriting authentication systems.	[6] [10] [11]	Past research on forgeries either targeted signatures or freeform writing patterns using a tablet. Our work is the first study which investigates freeform handwriting authentication under attacks in which forgers mimic the wrist-movement patterns of the victim.

1) Signature-Based Authentication: Most research on handwriting authentication involves the use of signatures. They are largely consistent over time and unique, making them an appealing form of verification. Although research on signature verification has been going on for decades, the field continues to advance by integrating new hardware and machine learning techniques.

One publication closely related to our own research is that of Levy *et al*. [6]. They used a Microsoft Band to authenticate users based on their signatures. Sensor data from the Microsoft Band was processed using the discrete cosine transform (DCT) and dynamic time warping (DTW) to extract feature vectors. Several classifiers including Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression (LR) were then applied to the feature set. They achieved an EER of less than 0.99% for random forgeries.

More recently, Tolosana *et al*. [7] demonstrated the use of a Siamese neural network architecture for signature authentication using data from a tablet. In the case of skilled forgeries, the Siamese Bidirectional Long Short-Term Memory architecture was able to achieve an EER of 5.50%. The authors also compared their results with that achieved through DTW. They found that although DTW performed better for random forgeries, the Siamese network outperformed when it came to skilled forgeries.

Other recent research has also added to the literature on signature authentication in various ways. Gomez-Barrero *et al*. [13] achieved an EER of 0.5% with DTW, decreasing the EER by over 50% by using the Sigma LogNormal model as a representation of stroke movements. Diaz *et al*. [14] tackled single reference learning for signature authentication. They achieved this by decomposing the signature using the Sigma LogNormal model and creating new training instances based on stroke and target modifications to the signature. Additionally, Houmani *et al*. [15] published a comprehensive benchmark study for researchers working on similar problems. Notably, the authors found that low quality (high entropy) signatures had a best performing EER 2% lower than for low entropy signatures.

The main difference between our work and the previously mentioned studies is that we use freeform handwriting, while their models were based on signatures. This is important because signatures are fundamentally different from freeform handwriting. A signature is somewhat like a password in that it is made up of predetermined components that the user has heavily practiced and built some level of consistency in repeating. Freeform handwriting on the other hand is unpredictable in terms of the content, so authentication must occur by recognizing the underlying unique patterns of the writer. Due to the lack of predetermined content, freeform writing is likely to have a higher cognitive load compared to written signatures because a user is actively concretizing their abstract thoughts on paper rather than writing from memory. The decision making process involved in selecting words and considering phrasing is likely to lead to more pauses and slight hesitations while writing. These behavioral quirks are fundamental to what differentiates freeform handwriting from signatures, separating our work from research on signature-based authentication.

Due to these inherent differences between signatures and freeform handwriting, past work on signature authentication sheds little light on freeform handwriting authentication. Next we explore related research which investigated freeform handwriting authentication.

2) Freeform Handwriting Authentication: The vast majority of sensor-oriented handwriting authentication leverages data gathered with smart pens that have embedded sensors, or tablets which measure movement and touches across their surface.

One of the papers most closely related to our own work is that of Zhang *et al*. [8]. In that research, the authors gathered data from over 100 users writing both Chinese and English in freeform writing scenarios. Their data was gathered from a WACOM tablet, which allowed them to measure when the user was writing and when they were not. By using pen-up periods as "imaginary strokes", they were able to combine real and imaginary strokes into sets of (x, y) coordinates that were fed into a bidirectional Long Short-Term Memory (BLSTM) layer. After using majority voting with multiple random samples, this

method achieved 100% accuracy for the English dataset and 99.46% accuracy for the Chinese dataset.

A number of other innovative studies more distantly related to our own work have explored freeform writing or motion-based authentication. For example, freeform doodling (drawings) was explored by Martinez-Diaz *et al*. [9] as a method of phone authentication. Users practiced drawings a small number of times before repeating them for the authentication process. This method achieved EERs of 3-8% on random forgeries and around 21% on skilled forgeries. A paper by Tian *et al*. [16] then investigated the use of a leap motion controller for in-air writing authentication by tracking the movement of a user's finger. This method achieved an EER of as low as 1.18%, or 3.11% for observation attacks (with 4 victims). These works relate to our work by virtue of evaluating user handwriting patterns, however there are several significant differences.

The major difference between past works on freeform handwriting authentication is that they use a different vector for data collection. Zhang *et al*. and Martinez-Diaz *et al*. relied upon tablets or phones to gather data and act as a writing surface. As a result of recording the location and time of contact on the surface, the data driving their authentication system contains the exact characters as they are written by the author. These precise details of the character shapes and relative locations provide a significantly different type of information about the writing process as compared to a wrist-worn device recording wrist movements. The device used in our work only measures the acceleration of the hand and does not provide detailed information on the shapes of the characters written by the user. Due to our method not using direct information on the spatial characteristics of the characters, our work tackles a different variant of the handwriting authentication problem as compared to past work involving freeform writing on tablets or phones.

In the next section we describe how our research differs from past work when evaluating handwriting authentication systems against non-zero effort threat models.

B. Non-Zero Effort Attacks on Writer Authentication Systems

By virtue of evaluating attacks on our freeform handwriting authentication system, our work is related to the study of impostor attacks against sensor-oriented handwriting authentication systems.

Past work has studied impersonation attacks against signatures, with most of the recent work involving tablets or mobile devices (e.g., see [14], [7], [13]). A significant area of focus now is the creation of impersonation resistant systems. One such example is the Siamese network architecture in [7] where the EER increase was shown to be limited to around 1% if trained using skilled and random forgeries. Overall, the attacks demonstrated a 1-7% increase in the EER due to skilled impersonation attacks as compared to random forgeries.

Perhaps more related to our own work is that of Levy *et al*. [6], where the authors also used a wearable device for authentication but they used signatures as the form of writing. The authors used a tablet to record the strokes of users writing their signature in order to facilitate an effective attack. The authors then allowed attackers to view these recordings and practice their imitations in real-time before conducting the imitation attack. This attack by Levy *et al*.involved an attacker impersonating static text (i.e., a signature) from a given user. They increased the EER of the authentication system from 0.99% to 2.63% after skilled forgeries were used. The wrist-based method devised by the authors also outperformed the state-of-the-art SUSIG model, with an EER of 2.63% versus 5.50%.

A much earlier work by Humm *et al*. [11] on the other hand conducted attacks on text-dependent and freeform writing authentication systems. In both cases they found that the attacker was able to significantly increase the EER of their target (from 4% to 13.7%) by copying the text that the target had written. Particularly notable about this attack experiment was that the attackers were limited to only a few minutes of practice per forgery, while the authors acknowledge that real attackers would likely practice for hours. However, the authors had participants write on a tablet for the experiment.

The above series of works have two common attributes that distinguish them from us: they either use touch-sensitive surfaces to record writing patterns or focus on signature authentication (some do both of these). As previously described in Section II-A2, these differences in the form of writing and nature of writing measurements imply that our work tackles a different variant of the handwriting authentication problem from that tackled in these papers. Our paper is to our knowledge the first work to investigate the question of a trained impostor spoofing freeform handwriting as captured from wrist movement patterns.

III. Handwriting Experiments

After receiving Institutional Review Board (IRB) approval; we conducted five experiments (E1, E2, E3, E4, and E5) to test the effectiveness of freeform writing for user authentication given a variety of conditions and threat models. These experiments broadly fell into two categories: *core authentication*, in which we gathered data for training and testing smartwatch-driven freeform handwriting authentication systems, and the *impostor experiments*, where we trained attackers to target specific users based on information they are given about the target. We first describe the operational details and devices used for data collection, we then describe the specific details of individual experiments and reasoning behind them.

A. Procedures and Gadgets

Our 53 participants were recruited from across campus. Most of them were undergraduate and graduate students. We collected accelerometer data from a LG Urbane 2 smartwatch during the writing process for all participants. The smartwatch was paired via Bluetooth with a Samsung Galaxy 6 and it collected data at a rate of around 100 Hz for linear acceleration[2] measurements across three axis (x, y, and z). An app on the phone had buttons for starting and stopping the experiment, or saving and clearing

[2]According to the Android API specifications, linear acceleration is the accelerometer measurements excluding gravity.

TABLE II
A NON-COMPREHENSIVE LIST OF E3 SCENARIO QUESTIONS.

Bloom's Level	Cognitive Load	Action
Remembering	1 (Low)	Tell about a group project experience.
Understanding	2	What technological advances are you most looking forward to over the next 30 years? Why?
Applying	3	Explain how you think society and the world would be different if a given emotion didn't exist.
Analyzing	4	What would you have done differently without your most inspirational experience?
Evaluating	5	What skills should students learn in high school that they don't already learn? Why?
Creating	6 (High)	Imagine achieving your goals. Plan out the steps it will require to achieve those goals.

the data. After pressing the start button on the phone app, the smartwatch app collected data and sent it in batches to the phone. The phone received the data and recorded it for the creation of the authentication system. A typical experimental session took an hour to complete.

We also collected video of the writing process for 30 participants. The video was used to help attackers practice writing like their target. Participants were informed as to the procedures and activities, and all users for whom video of their hand was recorded were informed that this would be occurring. During the experiments, the participants were asked to be seated at a table with a piece of paper and pen for writing. They remained seated throughout the duration of each experiment. Participants were asked to avoid using the hand with a smartwatch to mess with their hair or clothing during the sessions. We also asked that participants keep their hand steady while pausing and flip the piece of paper over with their other hand.

B. Data Collection for Core Authentication

The three writing experiments (E1, E2, and E3) were dedicated to the core authentication. Each of these experiments were conducted twice, once on each of two separate days. This allowed us to incorporate some of the variance that could occur in a user's behavior over the course of multiple days due to changes in mood or energy level. Due to the distance between the smartwatch and writing tool, we investigated whether video would act as a better source of information for impostors. For the last 30 users (of the 53 total) we also collected video of their hand during the writing sessions. Next we describe the experiments E1, E2, and E3 in detail.

1) Experiment E1: We had all users copy the exact same speech excerpt during all sessions from a printout to a piece of lined paper. We did this to investigate the possibility of algorithms being able to differentiate between users entirely based on writing patterns, even when the content is the same (e.g., some note taking scenarios). For this experiment we had users write an excerpt from Jeff Bezos's 2010 graduation speech at Princeton.

2) Experiment E2: Each user copied a unique piece of text during each session. Compared to experiment E1, this scenario provides a more realistic representation of how users might be differentiated between based on copying text which is unique. In practice, users of a handwriting authentication system are unlikely to be copying the same text. The text that users copied came from a wide variety of speeches, most of which were selected from the website American Rhetoric and its list of the top 100 speeches.

3) Experiment E3: This experiment involved each user writing answers to a unique set of open-ended questions provided on each day of data collection. We developed this experiment to evaluate how a freeform handwriting authentication system handles written tasks that require thought and consideration, or even pausing to think.

To ensure we were accurately modeling the types of scenarios in which a user might engage in considerable thought while writing, we used Bloom's Taxonomy (see overview [17]), a common framework in education for assessing learning and cognitive load. Bloom's Taxonomy ranks types of cognitive tasks based on their complexity and demonstrated level of understanding. It has been used in a diverse set of applications, such as the evaluation of sensor-based learning prototypes, evaluating an automatic tutoring system, or in designing experiments for improving e-learning [18], [19], [20]. In our research we used Bloom's Taxonomy to design questions that would catalyze varying levels of cognitive load in our participants. Table II shows example questions from our experiment and their matching cognitive loads according to the blooms taxonomy. Each question had multiple parts addressing different Bloom's Taxonomy levels, this meant that participants had multiple opportunities to be exposed to each level of the taxonomy.

C. Data Collection for Impostor Experiments

Experiments E4 and E5 were dedicated to impostor threat evaluation. They were designed to test how different types of imitation attacks might affect the handwriting authentication system. Below we describe these two experiments (additional experimental details can be found in Section VI-B).

1) Experiment E4: We allowed attackers to view and practice against training samples of their target's handwriting before trying to impersonate them. This attack is designed to be similar to the situation in which an attacker comes across a sample of their target's writing or steals it for practice. The attack enabled us to explore if this type of information about the target would allow an attacker to breach the system. Impostors were given samples of the victim's handwriting for at least a week in advance of the experiment. We asked all impostors to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2019.2912401, IEEE Transactions on Biometrics, Behavior, and Identity Science

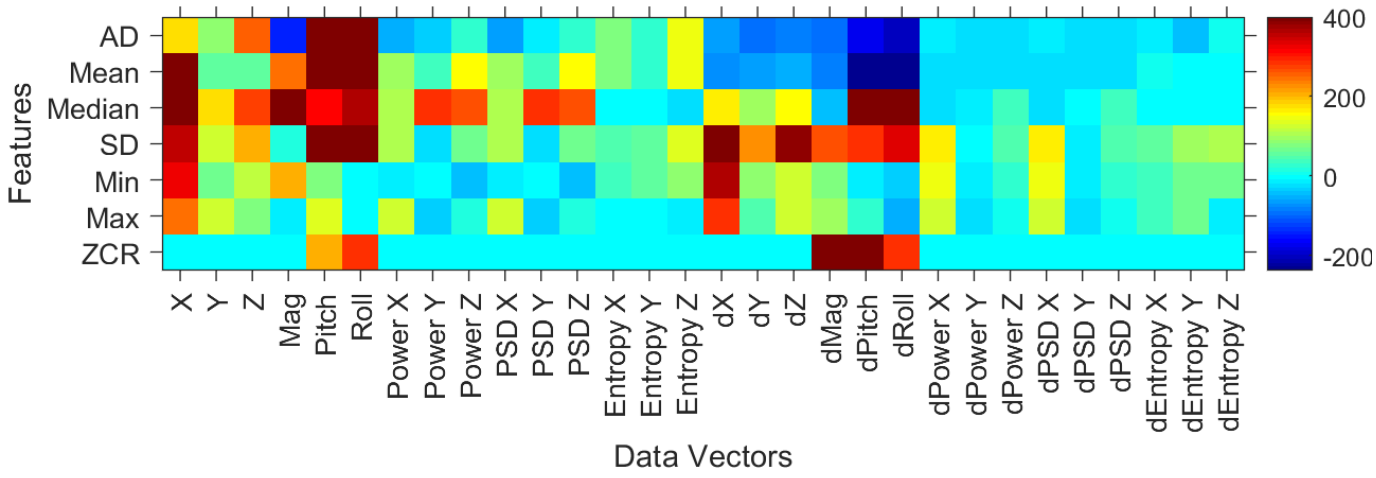


Fig. 1. The ReliefF values for each feature, the x-axis is the data vector and the y-axis is the mathematical operation applied to the data vector to get the feature. SD stands for standard deviation, AD represents absolute difference. Higher positive scores designate features that are more likely to be predictive. We made 400 the maximum value on the colorbar to make it easier to distinguish between different features.

practice writing like the target based on the physical writing alone, before the session in which we gathered the data. At the end of the training, each impostor was required to write at least one full page worth of practice, which we inspected before the experiment. During the experiment, attackers were allowed to have the copy of the text in front of them as an example.

2) Experiment E5: For this experiment, attackers were allowed to view video of their target writing during the training session. Unlike E4, which focuses on the visible features of the written text, this attack captures the scenario in which an attacker additionally mimics the motion dynamics of the victim's wrist. Most attackers practiced an additional 30 minutes or more for each of their targets after receiving the video (which was not initially given to them).

IV. Feature Extraction and Authentication Paradigms

As described in our contributions, we explored two design paradigms for a freeform handwriting authentication system. One based on human-engineered features (i.e. feature engineering), the other based on feature learning. In this section we describe how we implemented these two approaches and their associated data processing procedures.

A. Feature Engineering-Based Authentication

To drive our feature engineering based authentication system we extracted a set of 210 features from the accelerometer data. In this section we describe the process of extracting these features, feature selection, and the associated classification algorithm used in this authentication approach.

1) Data Processing and Feature Extraction: To reduce the influence of outliers or inconsistencies, we used a third-order lowpass butterworth filter to smooth the data. This filter has been successfully used in biometrics research for this purpose (e.g., see [21], [22]). Before extracting the feature set we randomly selected 5 second windows of data, with the number of windows equal to the number of samples in a session divided by the window size.

Our feature extraction was a 2-step process as described next. Step-1: We computed 15 data vectors from the accelerometer and computed 15 additional vectors that express the derivative (or rate of change) of the first 15 vectors. We have used the rates of change in our work since they have been previously shown to originate highly discriminative features in related biometric modalities (e.g., see [23]). The x-axis of Fig. 1 shows the names of these 30 data vectors[3]. Data vectors preceded with d are the derivative vectors (e.g., dX is the derivative of vector X). Note that except for the data vectors X, Y, and Z, which contain raw sensor readings, all other data vectors capture various well-known frequency and time domain attributes of a time series signal that are often used for feature computation;

Step-2: From each of the 30 data vectors, we then computed 7 different features or metrics (namely, mean, standard deviation, absolute difference, min, and max values, median, and the zero crossing rate), creating a total of 210 (or 15 x 2 x 7) features. These different features are captured on the Y-axis of Fig. 1 (the acronymns on the Y-axis have the following meanings: ZCR — the zero-crossing rate, SD — standard deviation, AD — absolute difference, Min — the minimum, and Max — the maximum).

Having extracted the features, we then split our data into three datasets, namely a training set which comprised of 60% of each user's data collected in the first session, the validation set which which comprised of 40% of each user's data collected in the first session and the test set which comprised of all data collected in the second session.

2) Feature Ranking and Selection: After extracting 210 features from the 30 data vectors, we used the Scikit-Learn ([24])

[3]The acronym Mag stands for the magnitude ($\sqrt{x^2+y^2+z^2}$) while the PSD is the Power Spectral Density

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2019.2912401, IEEE Transactions on Biometrics, Behavior, and Identity Science

StandardScaler with default settings for feature normalization. To cut down this large set of features to a compact and highly discriminative subset, we applied the ReliefF feature ranking method [25] and retained the top 30 features. It should be noted that scaling and feature ranking occurred based on the training data. We chose the top 30 features because a greater number didn't offer significant gains in performance. Fig. 1 shows the ranking of these 210 features in terms of a colormap based on the ReliefF values for each feature. The colorbar on the right side of the figure demonstrates how features are ranked. High numbers corresponding to colors like dark red are stronger features, while low or negative numbers represent features that aren't likely to contribute to strong system performance.

We found that the features could be invaluable in terms of the interpretability of the model. Consider pitch and roll, these two data vectors represent the motion of a user's wrist as they are writing. Different styles (e.g., fast or slow) of writing are likely to express themselves in unique ways through the rotation of the wrist during the writing process. We found that frequently if a single feature was high (or low) ranked for a data vector, other features extracted from that data vector would often show a similar trend.

3) User Authentication with Engineered Features and Classifier Ensemble: Based on the top 30 features selected based on the previous section, we explored a range of classifiers in the Scikit-Learn package [24]. We found that Logistic Regression (LR) and Gaussian Naïve Bayes (NB) outperformed other classifiers. We then found that combining these two classifiers through score-level fusion with the weighted sum rule (see [26]) and equal weights improved the results on a subset of the data. Therefore for our authentication system we used an ensemble of these two classifiers (hereafter referred to as LR + NB classifier).

During the training process we used a 2 to 1 ratio of impostor vs authentic samples. Different users had different numbers of authentic samples in the data depending on how long they took to complete the writing exercise. In general, most users had between 200 and 400 authentic samples.

B. Feature Learning-Based Authentication

1) Overview of Feature Learning Model: To learn features from the raw sensor data, we used a deep neural network with stacked Convolutional Neural Network (CNN) and BLSTM layers. Similar sensor signal fusing DNN architectures in recent works include those of [27] and [28]. Our network architecture (see Fig. 2) has two convolutional blocks and a BLSTM for each sensor axis. Each convolutional block is comprised of two 1D convolutional layers (kernel size of 3, with either 32 or 64 filters), batch normalization, max pooling (size of 2), and dropout (50%). The output of the convolutional blocks gets fed into a bidirectional LSTM (BLSTM) with 10 neurons in each direction, where the BLSTM only returns the result from the last neuron. Convolutional and BLSTM layers used the Rectified Linear Unit (ReLu) activation function [29]. The output of the BLSTM layer for each axis gets concatenated and dropout gets applied once more before classifying the user as either authentic or an impostor using an fully connected layer with a softmax activation. The loss is calculated with cross entropy and we used Adam [30] for the optimizer with default parameters.

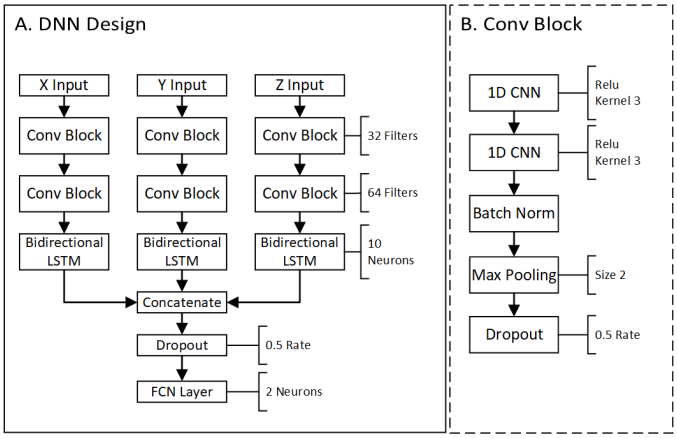


Fig. 2. On the left side of the figure (sub-figure A: DNN Design) is the overall architecture of the deep neural network (DNN) we used to classify users. The architecture has separate sub-networks which process each individual sensor direction (x, y, or z). Each sensor direction is processed separately with convolutional blocks (see sub-figure B: Conv Block) and a bidirectional Long Short-Term Memory (LSTM) layer before being concatenated and classified by a Fully Connected Layer (FCN). Further examination of sub-figure B shows that we use dropout and regularization to reduce overfitting, this was done in combination with data augmentation to allow the training of a deeper neural network.

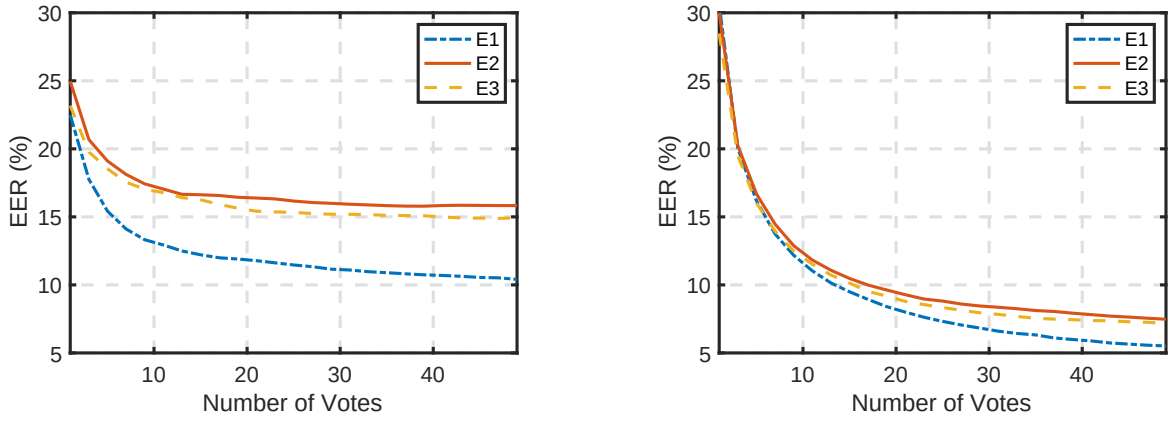
Due to training the DNN to differentiate impostors and an authentic user, we trained a DNN for every user. Calculating the EER meant using the probabilities outputted by the DNN to calculate the FAR and FRR. The probability thresholds for the EER calculation ranged from 0.0 to 1.0 using steps of 0.01. Note that the EER decision threshold for every user is selected using the validation data and reused on the test data.

2) Data Processing and Training: Given 3 data streams (x, y, and z axis), from the accelerometer we extracted windows of data to feed into the DNN. Using randomly selected windows of data from a given user, we created a matrix $X_{n\times 120\times 3}$ where 3 is the number of data streams, 120 is the number of samples from each data stream, and n is the number of data windows. This data was fed into the DNN in mini-batches of 4000 windows. We used this process for the train, test, and validation datasets.

To boost the ability of our system to identify attackers, we used the common trick of penalizing the dominant class [31]. Through experimentation we found that a 30 to 1 ratio of impostor data to real user data in the training set was the most effective. We then changed the learning rate to be biased against the dominant class in the training data, based on the ratio between the two classes. Test data for each user was a balanced random sample of authentic users and impostors. Training time for each experiment required over 72 hours, or around an hour and a half per user (including data processing). We used a computer with a Nvidia GTX 1080 Ti and twelve virtual cores for parallelized data processing.

3) Reducing Overfitting Through Data Augmentation: Deep neural networks require significant amounts of data to learn

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2019.2912401, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Feature Engineering - LR + NB

(b) Feature Learning - DNN

Fig. 3. Demonstrating the impact of majority voting on our two authentication systems. In both figures, the EER decreased rapidly over the first 10 votes for both systems, however, the DNN improves at a faster rate. Although the DNN began to stabilize around 30 to 40 votes, the EER for the LR + NB classifier only started to level out near 50. Overall the DNN performed better than the LR + NB classifier, and out of our experimental scenarios E1 performed the best and E2 the worst. We go into these results further in Section V-B.

generalized models for their given objective. Lack of enough data tends to cause overfitting and poor generalization [32]. Recent research (e.g., see [32], [33], [34]) has shown that data augmentation can be used to reduce overfitting and allow the training of deeper networks that can learn more discriminative features. For example, in image recognition, data augmentation is often used to extend a smaller set of training images by generating additional images that capture variations in attributes such as lighting or rotation (e.g., see [33]). For applications using time-series sensor data (e.g., accelerometer measurements) with DNNs, data augmentation is increasingly being used to capture variations in motion dynamics that might not be represented in a training set. For example, Um *et al*. [32] applied a wide variety of augmentation operations (e.g., rotations, scaling, permutation, etc) to add small variations to the accelerometer data captured from a small number of Parkinson's disease patients. These operations captured potential attributes of patient behavior that may not exist in the initial data and improved their accuracy of Parkinson's Disease state classification from 77.54% to 86.88%. In this paper we also applied data augmentation to improve the ability of our DNN to generalize effectively.

A key component of properly applying data augmentation is to consider the properties of the data being used to train the DNN. Augmentation should not alter the outcome or true label of the data. We investigated several methods of data augmentation before deciding to use a combination of scaling and rotation. Rotation helps simulate different orientations of the watch during the writing process. Scaling increases or decreases the magnitude of acceleration per-axis, as users may write at a different pace across sessions.

We used the implementation by [32] for our data augmentation, including the default parameters for the scaling and rotation operations. Given a window of data, we needed to decide whether to apply scaling, rotation, both augmentations, or no augmentations. We used a 25% probability for each of these four options. The augmentation operation is conducted as follows: (1) We randomly generate an index between 0 and the maximum index that depends on the amount of data available for a given user, (2) we use this index as the starting point for a window of data selected from the user's raw time series measurements, (3) we make an augmentation decision on this window as previously discussed, (4) we conduct this process n times, where n is selected such that the final augmented dataset is approximately four times the size of the original dataset. *Note that augmentation was only applied to the training dataset, not the validation or test set.* To stay consistent with our overall authentication system design process we also tried applying data augmentation to the data used for feature engineering. We did not find that feature engineering saw increased generalization capability and increased accuracy from this step. From this point forward in the paper it should be assumed that the DNN is using data augmentation and LR + NB classifier is not.

V. Experimental Results

In this section we describe the performance of our core authentication system for both the feature learning and the feature engineering paradigms. Our design is motivated by the fact that the ideal freeform handwriting authentication system should authenticate a user using as little data as possible. The advantage of using very little data is that an impostor would be detected quickly.

In the following sections we first present our authentication system performance evaluation based on single windows of data. We then analyze the impact of using majority voting on multiple windows to boost system performance. All results in this section are based on the zero-effort threat model in which adversaries of the authentication system make no effort to forge their data.

A. Mean Authentication System Performance - Single Vote

Here we present the error rates across the population using a single vote (or single window). Fig. 3a and Fig. 3b on the

extreme left show these results. We can see that the results are very poor for a single vote of the classifiers, as the EER is about 28-30% for the DNN and around 22.5-25% for the LR + NB classifier. These results indicate that individual windows of handwriting data do not provide enough discriminative information to rigorously characterize an individual's wrist movement patterns during writing.

B. Mean Authentication System Performance - Window Fusion

To improve on the results seen for the individual data windows, we used a fusion approach based on majority voting (i.e. a final decision is made based on the majority of decisions for multiple voting windows). We conducted this majority voting evaluation for odd numbers of windows between 1 and 49 (inclusive).

Our results for this procedure when both the LR + NB classifier and the DNN are shown in Fig. 3. This figure shows that for both the LR + NB classifier and the DNN, this approach is highly effective at reducing the EER beyond the error rates that were obtained for a single window of data. While the EER continued to decrease for the DNN until around 50 votes, it stabilized for the LR + NB classifier after around 30 votes. After just a few votes, E1, E2, and E3 showed a consistent pattern in that E1 performed best and E2 performed worst. Table III shows our best results for majority voting using the 49 vote configuration. For all three experimental scenarios, the DNN outperformed the LR + NB classifier.

A possible reason for the difference in performance between E1 and the other scenarios is the consistency of the content being written. E1 was specifically designed such that users were writing the same content during each session, making E1 somewhat analogous to a signature (albeit with a greater amount of text). This commonality of character sequences between the training and testing sets will have enabled the classifiers to more easily recognize certain user-centric patterns associated with these sequences, which in turn likely made each user more differentiable from other users by the authentication system. On the other hand, E2 and E3 were designed to have the user write different content in each session, rendering the classifiers less able to leverage content-specific patterns to beef up classification performance. It is noteworthy that the poor performance of the LR + NB classifer relative to the DNN might be because of the specific features fed to the LR + NB classifier, as opposed to being due to its classification mechanism. The DNN is able to mitigate this additional feature-centric uncertainty by learning its own feature abstractions.

To put our findings in the context of the state-of-the-art, we note that the closest related research on handwriting authentication is signature-based and (or) tablet-based. Recent research in these domains has reported error rates of between 0.5% and 4% (e.g., see [7], [6]). Our DNN results thus approach the upper bounds of these results, while our LR + NB results are much higher than these results. As earlier explained however (see Section II), the input to our authentication system is much less precise than systems for which writing patterns are captured on a tablet surface. Also, relative to signatures which are just a few heavily practiced characters, our approach is susceptible to significant amounts of variation as a writer composes freeform writing text.

TABLE III
COMPARISON OF THE EERS FOR OUR AUTHENTICATION SYSTEMS.

		Experimental Scenarios EER (%)		
Classifier	**Feature Paradigm**	**E1**	**E2**	**E3**
DNN	Feature-Learning	5.51%	7.47%	7.18%
LR + NB	Feature-Engineering	10.40%	15.82%	14.92%

VI. Demonstrating State-of-the-Art Attacks on Wearables-Driven Handwriting Authentication Systems

A. Threat Model

Until this point, our research has been evaluated based on the zero-effort threat model. As mentioned in Section III, we developed two attacks, E4 and E5, to evaluate how our authentication system handles non-zero-effort impostors.

In the E4 attack scenario, we had impostors practice writing like the target by viewing, tracing, and imitating a writing sample from the victim (see Section III-C1). This was done to imitate the way in which an attacker is likely to come across writing samples from their target. People in a normal office, home, or university environment are likely to be relatively careless with written documents. Unwanted handwritten documents often end up in a waste paper bin or a normal trashcan, rather than a shredder or incinerator. If the attacker could be highly successful with an imitation attack based on this readily available source material, it would be cause for concern.

Impostors in E5 were given video of their target writing (see Section III-C1), after the E4 experiment completed. By giving attackers video of the victim writing, we allowed them to examine the intricacies of the wrist motion dynamics of their target during the writing process. Realistically, given the ubiquitous nature of recording devices in phones and other hardware, an attacker can likely gain access to video of their target writing; particularly in school environments where handwriting in open environments is a commonplace occurrence. If wearables-driven freeform handwriting authentication were to become mainstream, there would be a market for this type of video and the system would need to be built to be resistant to such attacks.

A noteworthy aspect of impersonation attacks in a handwriting authentication scenario is that the attacker and user may collaborate towards a common goal (e.g., during an exam). They could work together, with the target providing written and video samples for the attacker to use as practice. This sets handwriting authentication apart from other biometric authentication systems, because rarely is the target incentivized to aid their attacker. This further motivates the need to evaluate handwriting authentication against sophisticated attacks. The results of the extensive threat assessment are discussed in greater detail within the following subsection.

TABLE IV
IMPACT OF THE ATTACKS ON DIFFERENT WRITING SCENARIOS AND CLASSIFIERS.

Classifier	Pre-Attack FAR (%)	Post-Attack FAR (%)	Change in FAR (%)	Handwriting Experiment	Attack Strategy
LR + NB	6.49	13.66	7.17	E1	E4
LR + NB	10.81	16.77	5.96	E3	E4
LR + NB	6.49	10.84	4.35	E1	E5
LR + NB	10.81	13.82	3.01	E3	E5
DNN	3.31	2.50	-0.81	E1	E4
DNN	6.87	11.06	4.19	E3	E4
DNN	3.31	8.7	5.39	E1	E5
DNN	6.87	4.03	-2.84	E3	E5

B. Evaluating the Impact of Determined Attackers

1) Impostor Recruitment: One of the most important aspects of impostor recruitment, was the process of selecting participants who would take the task seriously and make an effort to practice in their free time. In order to select participants who fit this profile, our recruitment went through three phases. First, we asked for volunteers during the initial data collection. Second, we emailed the volunteers several weeks later once the attack experiments were about to begin. In general, if a student emailed us back, this was a basic preliminary sign of interest. Third, we requested that they rate themselves on a scale of 1 to 5 (see Likert Scale [35]), representing how motivated they were to practice and take part in the impostor attack phase of the research. These multiple phases of the filtering process were designed to eliminate participants who were unmotivated.

Given this subset of motivated participants, we further incentivized the impostors to make their best effort by promising a reward of 20 dollars to the attackers who perform best in each of several categories. A single impostor was able to win multiple categories, so there was an additional incentive to make an effort in every category. To ensure attackers were able to do their best, we explained how our authentication system worked and the general way in which it authenticates users. Secure authentication systems should not rely on being a black box, they should remain secure even if the attacker understands how they work.

2) Victim Selection: To get a representative subset of the different behavioral patterns in our population of users, we selected 27 victims for our attackers to target. The victims were selected randomly from among the larger set of users for whom we had video data. Because users have a range of authentication template strengths (e.g., weak vs strong), we tried to evenly distribute target template strengths among attackers. We divided victims into three groups (low, medium, and high) based on their EER. A low EER indicates that the user's template is effective at distinguishing that user from others in the population. While a high EER indicates that their writing pattern is inconsistent or prone to misclassification with other users.

These authentication template strength categories were then used in distributing targets to our attackers. Due to the hours of practice required to prepare for impersonating other users, we decided to not have impostors target every victim. Instead, they were given an equal number of targets from every category of authentication template strength. This allowed them to rigorously practice against a smaller number of targets.

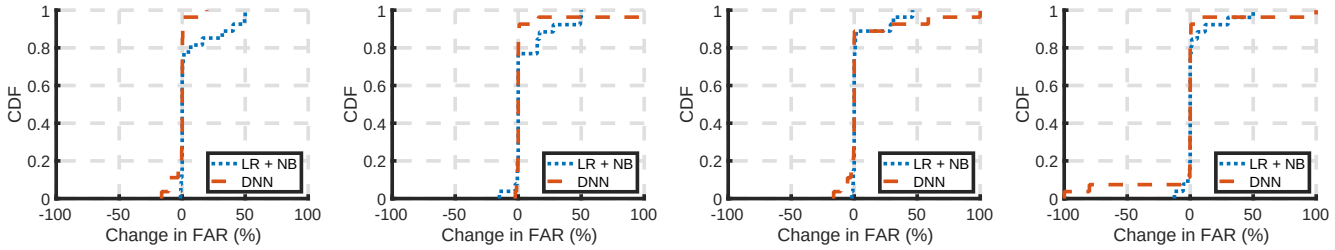
3) Classification Mechanisms for Evaluating Attacks: These attacks were implemented on the baseline authentication system described in Sections IV-A and IV-B. The only difference was that during the testing phase, impostor samples were selected from trained attackers. By doing this we were able to measure the impact of the imitation attacks in comparison to the zero-effort attacks. The impact of our attack is quantified using the False Acceptance Rate (FAR), which is the number of successful impostor attempts divided by the total number of attempts. We chose to use this metric because it directly measures the success rate of impostors. Because a decision threshold must be used when calculating the FAR, we used the same threshold for the attack (per user) as is used during the baseline experiments when calculating the EER for each model and experimental dataset. This is a close approximation to how the authentication system would work when deployed, where it would likely have different pre-configured thresholds per user.

4) Population-Level Impersonation Attack Results: Table IV shows the mean FAR of the attack on the entire user population across our different experimental scenarios. We attacked the E1 and E3 scenarios because they represent the two extremes of possible applications. E1 is highly constrained, such as if a system used a static prompt for authentication, while E3 closely imitates real-world scenarios (e.g., essay writing). These results allowed us to draw several general conclusions.

(i) In all our experimental scenarios, we found that the LR + NB classifier saw increased EERs due to the spoof attack. The DNN however had mixed results as some cases saw decrements in EER while others saw increments in EER due to the spoof attack. This trait suggests that the LR + NB classifier might be more prone to spoof attacks than the DNN. However, it should be noted that the comparatively lower attack resistance depicted by the ensemble might also be due to dynamics of the particular set of human-engineered features fed to it as opposed to the classifier mechanism.

(ii) When attackers had access to both the victim's static

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2019.2912401, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) E4 Attack on E1 (b) E4 Attack on E3 (c) E5 Attack on E1 (d) E5 Attack on E3

Fig. 4. Illustration of user-level impact of the spoof attacks. The figure shows Cumulative Density Functions (CDFs) of the change in False Accept Rate (FAR) caused by Attacks E4 and E5 relative to the zero-effort attacks. The figure reveals that a small number of users see a disproportionately high increase in FAR for both attacks on writing scenarios E1 and E3.

written text and a video of the victim's writing (i.e., the E5 attack), writing scenario E1 always saw increased EER while writing scenario E3 had mixed results. This result suggests that writing scenario E1 might be more vulnerable to spoof attacks than writing scenario E3. A possible reason behind this is that E1 is based on a fixed text that is synonymous to a signature while E3 is a purely freeform writing setting that captures the earlier discussed cognitive dynamics that might be more difficult for the attacker to copy and spoof.

(iii) Independent of the specific writing scenario and classification mechanism, the results depict no significant difference in impact between attacks E4 and E5. In particular, each of the attacks E4 and E5 caused EER increments that were on average of comparable magnitude across three of the four experiments (both attacks saw reduced EER in the fourth experiment). This trait seems counter intuitive as one would expect attack E5 to always outperform E4 due to the detailed information accessed through watching the target's video in attack E5. Potential reasons behind this include variations in attacker forgery expertise, variations in victim writing patterns and even the the sheer difficulty of mimicking wrist movement patterns exhibited in a freeform writing which might have meant that the extra information available under experiment E5 did not provide that much leverage to the attackers relative to experiment E4.

One of the issues with the population-level results discussed above (results in Table IV is that it is difficult to judge how the user-level patterns in the results influenced the overall mean. In the next section we show how the user-level results provide more clarity.

5) User-Level Impersonation Attack Results: Fig. 4 represents the user-level results after the spoof attack. The figure shows the CDF of the FAR changes (FAR after the attack minus FAR before the attack) per target across all attack experiments. The following points summarize the main trends depicted in Figure 4.

(i) A very small proportion of users are badly affected by the attack – By observing the uppermost portions of each sub-figure in Figure 4, one can see that a very small portion of the target users were significantly vulnerable to this attack. Although the vulnerable users experienced a large increase in their FAR, around 80% of users for E4 and 85% for E5 had zero change or a decrease in their FAR. This means that the increases in FAR due to the attack were largely caused by a small population of users, while most experienced a decrease in FAR (or zero change). The fact that the attack only impacts a small portion of users suggests that a targeted solution (e.g., a failure to enroll policy) might be able to mitigate the attack impact.

(ii) LR + NB versus DNN resistance to impersonation attacks – Observe that the DNN curve is to the left of the LR + NB classifier curve in three out of four sub-figures (see 4a, 4b, and 4d). This explains the trend earlier described in Section VI-B4 — i.e., the higher FARs seen with the LR + NB classifier were because it had a larger proportion of users who experienced a significant increase in their FAR after the attack as compared to the DNN.

Overall these experiments indicate that the performance of these types of attacks may be dependent on the type of classifier, features, writing scenario, and specific threat model. With the careful design of a wearables-driven freeform handwriting authentication system it seems possible to design a system that is robust against the type of attacks described in this section. This is likely due to the subtle wrist-movement dynamics which can distinguish between different users.

VII. Discussion and Conclusions

In this paper we successfully demonstrate a freeform handwriting authentication system using wearable devices. As part of this we have explored multiple designs (feature engineering and feature learning) and majority voting for the authentication system. We also conducted sophisticated attacks against it to check if it would be resistant to trained impersonators with significant amounts of information about their target's writing patterns. As part of this work, we share insights into the relative strengths and weaknesses of such a system, which might aid future researchers or applications of our work.

Practical Usage and Deployment: In practice deploying this technology would be straight forward because the system is built to work with existing wearables technology using common frameworks, such as Android. For user experience, smartwatches are cheap and easy to use with respect to our authentication system. A user would simply provide a sample of writing data, a template would be created for them, and they would be able to authenticate against that template by

using an app on their smartwatch. One example of where this technology could be applied is in education, providing a new way to discourage cheating at scale. This being said, there are ways in which this research can be improved upon in future work.

Potential as a Unimodal System: Although promising, our system EER compared to some biometric authentication systems based on physical traits is above acceptable levels for a unimodal authentication system. This means that the greatest potential for our system lies in being part of a multimodal authentication system, whether using a password or other biometric traits.

Continuous Biometrics and Fusion Approaches: This authentication system could be used either as part of a continuous authentication system where the user is authenticated initially with several biometric traits and then continuously validated during the session using their handwriting. Or, the system could be fused with another biometric to make it stronger overall and more resistant to impersonation attacks by forcing the attacker to impersonate several biometric traits simultaneously.

By achieving an EER of nearly 7% for the E3 handwriting scenario, we demonstrate that handwriting authentication has significant potential, even when the cognitive load of the given task causes the user to pause while writing. We also show that the DNN achieves an average FAR of 6.57% after the attacks, demonstrating resilience to impersonation attempts, an important feature for deployed authentication systems. This research helps advance the state-of-the-art for freeform handwriting authentication. Our future work will focus on improving the authentication results further, exploring appropriate multimodal pairings for robust resistance to attacks, and reducing the amount of data required for accurate authentication results.

ACKNOWLEDGMENT

This research was supported by National Science Foundation Award Number: 1527795.

REFERENCES

[1] D. Impedovo, G. Pirlo, and M. Russo, "Recent advances in offline signature identification," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 639–642.

[2] A. Bensefia and T. Paquet, "Writer verification based on a single handwriting word samples," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, p. 34, 2016.

[3] A. Hamadene and Y. Chibani, "One-class writer-independent offline signature verification using feature dissimilarity thresholding," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1226–1238, June 2016.

[4] "Wearables-driven handwriting dataset," https://github.com/isaac-gs/smartwatch-handwriting-dataset.

[5] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, "Offline handwritten signature verificationliterature review," in *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*. IEEE, 2017, pp. 1–8.

[6] A. Levy, B. Nassi, Y. Elovici, and E. Shmueli, "Handwritten signature verification using wrist-worn devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 119, 2018.

[7] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Exploring recurrent neural networks for on-line handwritten signature biometrics," *IEEE Access*, vol. 6, no. 5128-5138, pp. 1–7, 2018.

[8] X.-Y. Zhang, G.-S. Xie, C.-L. Liu, and Y. Bengio, "End-to-end online writer identification with recurrent neural network," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 2, pp. 285–292, April 2017.

[9] M. Martinez-Diaz, J. Fierrez, and J. Galbally, "Graphical password-based user authentication with free-form doodles," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 607–614, 2016.

[10] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: an evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[11] A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication," *IEEE Trans. on Sys., Man, and Cyber.-Part A: Sys. and Hum.*, vol. 39, no. 1, pp. 25–35, Jan 2009.

[12] I. Griswold-Steiner, R. Matovu, and A. Serwadda, "Handwriting watcher: A mechanism for smartwatch-driven handwriting authentication," in *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, ser. IJCB 2017. Denver, CO, USA: IEEE, 2017, pp. 216–224.

[13] M. Gomez-Barrero, J. Galbally, J. Fierrez, J. Ortega-Garcia, and R. Plamondon, "Enhanced on-line signature verification based on skilled forgery detection using sigma-lognormal features," in *Biometrics (ICB), 2015 International Conference on*. IEEE, 2015, pp. 501–506.

[14] M. Diaz, A. Fischer, M. A. Ferrer, and R. Plamondon, "Dynamic signature verification system based on one real signature," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 228–239, 2018.

[15] N. Houmani, A. Mayoue, S. Garcia-Salicetti, B. Dorizzi, M. I. Khalil, M. N. Moustafa, H. Abbas, D. Muramatsu, B. Yanikoglu, A. Kholmatov *et al.*, "Biosecure signature evaluation campaign (bsec'2009): Evaluating online signature algorithms depending on the quality of signatures," *Pattern Recognition*, vol. 45, no. 3, pp. 993–1003, 2012.

[16] J. Tian, Y. Cao, W. Xu, and S. Wang, "Challenge-response authentication using in-air handwriting style verification," *IEEE Transactions on Dependable and Secure Computing*, no. 1, pp. 1–1, 2017.

[17] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory into practice*, vol. 41, no. 4, pp. 212–218, 2002.

[18] J. Schneider, D. Börner, P. Van Rosmalen, and M. Specht, "Augmenting the senses: a review on sensor-based learning support," *Sensors*, vol. 15, no. 2, pp. 4097–4133, 2015.

[19] N. K. Person, "Autotutor improves deep learning of computer literacy: Is it the dialog or the talking head?" *Artificial intelligence in education: Shaping the future of learning through intelligent technologies*, vol. 97, p. 47, 2003.

[20] L.-A. Ho and T.-H. Kuo, "How can one amplify the effect of e-learning? an examination of high-tech employees computer attitude and flow experience," *Computers in Human Behavior*, vol. 26, no. 1, pp. 23–31, 2010.

[21] S. A. Fattah, A. S. M. M. Jameel, R. Goswami, S. K. Saha, N. Syed, S. Akter, and C. Shahnaz, "An approach for human identification based on time and frequency domain features extracted from ecg signals," in *TENCON 2011-2011 IEEE Region 10 Conference*. IEEE, 2011, pp. 259–263.

[22] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[23] M. Martinez-Diaz, J. Fierrez, R. P. Krish, and J. Galbally, "Mobile signature verification: Feature robustness and performance comparison," *IET Biometrics*, vol. 3, no. 4, pp. 267–277, 2014.

[24] 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.

[25] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, Jan 1997.

[26] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern recognition letters*, vol. 24, no. 13, pp. 2115–2125, 2003.

[27] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 351–360.

[28] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning framework for eeg seizure detection," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[29] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, p. 947, 2000.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in nonstandard situations," *Machine learning*, vol. 46, no. 1-3, pp. 191–202, 2002.
[32] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for parkinsons disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 216–220.
[33] G. Urban, K. M. Bache, D. Phan, A. Sobrino, A. K. Shmakov, S. J. Hachey, C. Hughes, and P. Baldi, "Deep learning for drug discovery and cancer research: Automated analysis of vascularization images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.
[34] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.
[35] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

Isaac Griswold-Steiner graduated from Texas Tech University (TTU) in December of 2017 with Highest Honors from the Honors College (Summa Cum Laude). He now works at Microsoft as a Software Engineer.

Richard Matovu is currently a PhD student and a Graduate Teaching Assistant in the Department of Computer Science at Texas Tech University. He has been working on research projects focusing on privacy and security in mobile and wearable devices. His research interests include data mining, applied machine learning, biometrics and mobile security.

Dr. Abdul Serwadda is an Assistant Professor of Computer Science at Texas Tech University. He got his MS (in CS and Math) and PhD from Louisiana Tech University before doing postdocs at Louisiana Tech and Syracuse University. His research broadly designs learning algorithms for end-user security problems. Recent application domains for his research include mobile and wearable security, user behavior modeling and biometrics.