IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 5, 2018, accepted December 30, 2018, date of publication January 10, 2019, date of current version March 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891603

Improving Reliability: User Authentication on Smartphones Using Keystroke Biometrics

YUHUA WANG[1], CHUNHUA WU[1], KANGFENG ZHENG[1], AND XIUJUAN WANG[2]
[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Beijing University of Technology, Beijing 100124, China

Corresponding author: Chunhua Wu (wuchunhua@bupt.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0802703, and in part by the Science Foundation of China under Grant 61602052.

ABSTRACT Keystroke biometrics is a well-investigated dynamic behavioral methodology that utilizes the unique behavioral patterns of users to verify their identity when tapping keys. However, the performance of keystroke biometrics is unreliable due to its high error rate and low robustness. In this paper, we propose differential evolution and adversarial noise-based user authentication (DEANUA), which is a verification scheme for enhancing reliability by reducing the error rate and improving robustness. We investigate the current mainstream features and build a more comprehensive feature set that composed of 146 features. Then, we use a differential evolution method to select an optimized feature set. With the support vector regression method on this feature set, we achieve an equal error rate (EER) of 0.12660% and also a 31.25% energy consumption reduction rate. In this paper, the model is trained with the training samples collected from one situation, but the model is used in various situations. Thus, the robustness of the model is inadequate. We constructed the adversarial noise samples to simulate users' behavioral characteristics in different situational contexts. We use the adversarial noise samples to test the models in a strict experimental environment, which raises the EER by 83.59%, to 10.9299%. Then, we enhance the model with adversarial noise samples to obtain an EER of 8.70932%, which is a reduction of 20.32%.

INDEX TERMS Keystroke biometrics, touchscreen, authentication, behavioral recognition.

I. INTRODUCTION

Smartphones that store large amounts of private data are very popular, with 344.3 million sold worldwide in the first quarter of 2017 (1Q17) [1]–[3]. Normally, mobile devices are protected by static mechanisms like a personal identification number (PIN), fingerprint, and face recognition [4]–[7]. However, they are not entirely secure. Passwords can be guessed [8] or be attacked by side channels [9]. As for fingerprint authentication, the system can be spoofed by imitating the ridge and valley structure of the fingertip, which may have been generated from a latent fingerprint [10]. Facial recognition systems are also vulnerable to spoofing attacks, where a photo, video, or a 3D mask of a legal user's face can be utilized to gain illegitimate access to the system [11]. In order to overcome the weaknesses of static authentication systems, dynamic behavioral biometric techniques have been developed with the aim of being more difficult to imitate.

Keystroke biometrics has long been an active research topic in behavioral biometric techniques [12]–[17]. It can be classified into two groups: login authentication, and continuous authentication. This paper focuses on the login verification process. Research on virtual keystroke biometrics on touch screens first appeared in early 2008. Saevanee and Bhatarakosol [18] implemented a method, which aims to detect and authenticate the user on the basis of key hold-time, inter-key duration, and finger pressure. They achieved an Equal Error Rate (EER) of 1% on the basis of finger pressure. In 2014, Meng and Wong [19] implemented a scheme known as a touch dynamics-based authentication system. They used a Particle Swarm Optimization-Radial Basis Function Network (PSO-RBFN) with 21 features including pressure, speed, x-y coordinates, and gesture type, to achieve an EER of 2.46%. In 2016, Sitová *et al.* [20] collected data in both walking and sitting contexts and investigated why their features perform better in walking. They obtained an EER of 7.16% when walking and 10.05% when sitting. In 2017, Crawford and Ahmadzadeh [21] conducted a user study on the effects of movement while typing for authentication. They developed a two-stage approach to determine the user's behavioral context before classifying their typing behavior.

They first inferred the user's position with an Area Under Curve (AUC) of 90%; then they classified the user's typing pattern with an AUC of 93%.

Overall, most previous efforts have achieved an EER of around 0.5-10%. However, the European standard for access-control systems (EN-50133-1) specifies a false-alarm rate of less than 1%, with a miss-rate of no more than 0.001% [22]. So, there is a clear need for further research to improve accuracy. Also, many studies in keystroke biometrics have used the impostor's data as training samples for their classification. However, in the real world, most impostors will be new to the user's phone. Therefore, we need to determine how well the model will perform if the impostors are unknown during the training process.

Most work has studied keystroke authentication in one body position, but in daily life, the phone is used in various activities. Keystroke and touchscreen biometric models may be disturbed by different body motion conditions, e.g., lying, sitting, standing, and walking. Some innovative research has been done to conduct experiments in realistic contexts. As mentioned before, Sitová et al. analyzed their features in detail and achieved the best EER of 7.16%. They also measured the energy consumption of the accelerometer and gyroscope, sampled at 100Hz, 50Hz, 16Hz and 5Hz, finding that the energy overhead at 16Hz is 7.9%. If we need to collect accelerometer and gyroscope data all day for activity recognition, it will lead to a massive energy cost.

In Crawford et al.'s work, they determined the user's position with an AUC of 90% and the user's typing pattern with an AUC of 93%. As the accuracy of activity recognition is less than the accuracy of authentication, this means activity recognition may itself lead to more errors. Therefore, we propose a separate model for different circumstances. In addition, we use Adversarial Noise (AN) samples, which refer to crafting units such that they can indicate the user's actions in various body motion conditions. They are intended to cause misclassifications initially, and so to improve the training of our model. Similar experiments on PINs 1-1-1-1, 3-2-4-4 and 5-5-5-5 are conducted to allow comparison with Zheng *et al.* [23]'s work. Our contributions are summarized as follows.

- **Reducing the error rate:** We believe that a good feature set can improve the verification results. We investigate many relevant feature sets and build a more comprehensive one. Then we choose the typical feature optimization methods and conduct experiments to compare them. The feature ranking method only considers the relations between the feature and the result but ignores the interaction between features. Here, we collect as many as 146 features and test different feature selection methods including Pearson [24], Differential Evolution (DE) [25] and Binary Particle Swarm Optimization (BPSO) [26]. We use a DE feature selection method to find the best group of features. Then, after applying DE, a Support Vector Regression (SVR) method is used to achieve an EER of 0.12660% with impostors known to the model, and 6.35464% with unknown impostors.
- **Improving the robustness:** It is hard to collect and analyze data in all real situation contexts. However, we can analyze the distribution of data in different contexts and simulate samples. Generative Adversarial Networks (GANs) have been implemented using a system of two neural networks contesting with each other in a zero-sum game framework [27]. Inspired by GAN, we analyze each feature in all the samples from a user to find its robust distribution interval and build adversarial noise samples using the interval to imitate adversarial noise(AN) samples from various body positions. Then, we can generate adversarial samples to test the authentication model and enhance it. After adding adversarial noise samples to the validation set, the EER increased by 173.45% compared with the baseline method, whereas it only increased by 83.59% with Support Vector Regression (SVR). Hence, the SVR model has a stronger generalization ability than the baseline model with samples in different body motion conditions. Validation samples with adversarial noise cause misclassification and increase the error rate. Therefore, we add adversarial noise to the training process to augment training the model in different body motion conditions. This improves the robustness of the model, and the EER declined by 20.32% with SVR.

The paper is structured as follows. We review related research in Section 2. Then, we present a description of the Differential Evolution and Adversarial Noise-based User Authentication (DEANUA) scheme and evaluation methods in Section 3. Next, we introduce our dataset and methods in Section 4. In Section 5, we describe the authentication experiments and results. We draw conclusions in Section 6.

II. RELATED WORK

This section reviews related works in virtual keyboard keystroke dynamics arranged roughly in chronological and thematic order.

Maiorana *et al.* [28] focused on keystroke biometrics for user authentication on mobile devices. They proposed a statistical approach to guarantee verification rates when the number of enrollment acquisitions is low. Focusing on the use of alphabetical passwords, their database contained data acquired from forty users. They obtained an EER of 13.59%.

Trojahn and Ortmeier [29] developed a mixture of a mixture of handwriting-based and keystroke-based verification methods using capacitive displays. They proposed several new features like pressure during typing, fingertip size, and the physical characteristics of the mobile device. In their experiments, 18 users entered a specified sentence with 11 characters, ten times. They obtained their lowest EER of 1.13% with an RBFN method.

Zheng *et al.* [23] recruited 80 participants to an experiment using sensors such as an accelerometer, gyroscope, and the touch screen sensor on the smartphone to describe the user's

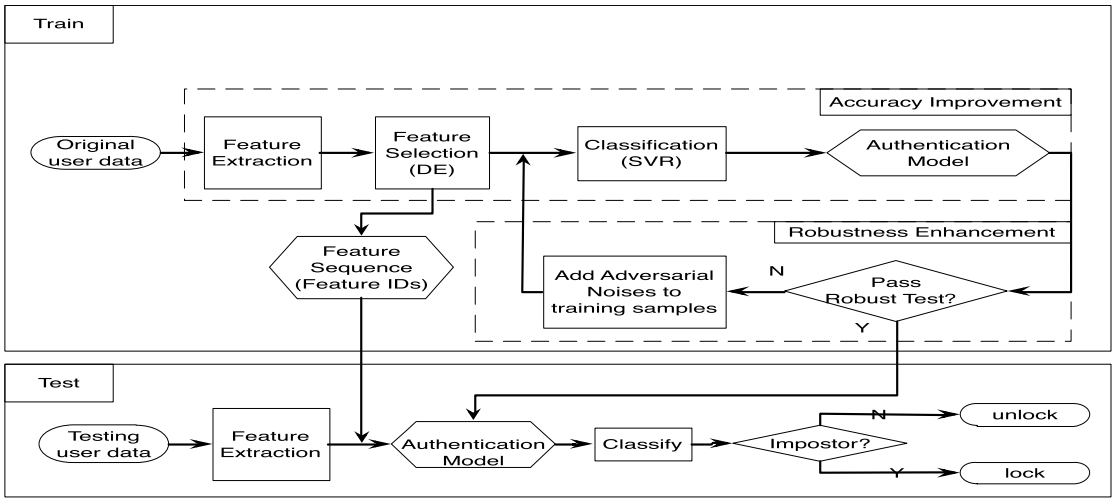


FIGURE 1. DEANUA Scheme.

unique behavioral patterns. Five different PINs: 3-2-4-4, 1-1-1-1, 5-5-5-5, 1-2-5-9-7-3-8-4 and 1-2-5-9-8-4-1-6 were used in their study. The authors used a one-class learning method based on nearest neighbor distance. The best EER was 3.58% for PIN 3-2-4-4. The other results were 7.34% for 5-5-5-5, 6.96% for 1-1-1-1, 4.55% for 1-2-5-9-7-3-8-4 and 4.45% for 1-2-5-9-8-4-1-6.

Giuffrida *et al.* [30] implemented an Android prototype system known as Unagi. Their implementation supported several feature extraction and detection algorithms for evaluation. They chose ''internet'' and ''satellite'' as passwords and focused on sensor-enhanced features. They also improved the accuracy of traditional keystroke dynamics, which previously had an EER greater than 7%, by two orders of magnitude.

Kambourakis *et al.* [31] proposed a methodology using only behavioral features, for example, distance and speed, corresponding to the way that the user interacts with the virtual keyboard. The authors designed two scenarios for both typical alphanumeric passwords and the variety of characters. Random forest(RF) [32], K-Nearest Neighbor(KNN) [33] and Multi-Layer Perceptron(MLP) [34] were used in the experiment. The best EER for the first and second scenarios was 26% and 13.6% respectively.

Morales *et al.* [35] presented user authentication using Keystroke Biometrics Ongoing Competition (KBOC). Their data included keystroke sequences from a legal user and impostors in a fixed text scenario. Thirty-one different algorithms were tested to find the one that gave the best accuracy and robustness. Their lowest EER was 5.32%. They also examined multisession variability by month of the year, and the accuracy degradation was less than 1% for probes.

Ataş [36] focused on hand tremor-based biometric recognition. They argued that hand tremor could help to authenticate users due to its unique characteristics. Fast Fourier transformation, discrete wavelet transformation, and 1-D local binary pattern methods were used to extract features from spatiotemporal hand tremor signals. They were able to exceed a 95% accuracy rate in classification tests.

The reliability of an authentication system is evaluated in terms of its error rate and robustness. However, the accuracy rate for access control systems has not yet reached the European standard, and the robustness of the model in different body motion conditions is rarely mentioned. We are concerned that a two-stage authentication for different body motion conditions may bring more errors with activity recognition and it will waste a lot of energy if we record the phone status continuously. From the accuracy perspective, we aim to refine the features and select the most useful set of features to obtain a lower EER than has been achieved in previous research. In addition, we also consider situations with unknown impostors and achieve energy use reduction. Concerning robustness, we add adversarial noise to test and strengthen the robustness of the model for all body motion conditions.

III. DESCRIPTION OF DEANUA SCHEME

The DEANUA scheme not only emphasizes the significance of accuracy but also pays attention to robustness. As shown in Figure 1, we split the DEANUA scheme into the training and test phases. In the training process, we first improve the accuracy by DE feature selection and pass the selected feature set (the selected feature IDs) on to the test section. Then we add AN samples to test the robustness of the model. If the model performs worse with the AN testing samples, we retrain the model with more AN samples to enhance the robustness of the model. After the robustness enhancement process, the model is retrained and passed on to the test section. In the testing process, the input data is from the user test data. After feature extraction and selection, the authentication model and threshold can help to classify the user into an impostor or legal user. Finally, the DEANUA scheme decides whether to unlock the mobile device or not.

A. ACCURACY IMPROVEMENT

In the training process, we focus on improving accuracy. As we believe that every acceleration that occurs between the user pressing and releasing a key contains useful information, we refine the features by adding more time features and every acceleration feature in every directional axis. It is necessary to choose a feature selection method that finds the most useful group of features. We prefer the DE feature selection method in this case, not only for its ability to take into account the relations between features but also its rapid processing speed and excellent overall performance.

Support Vector Regression (SVR) can determine the probability of a sample being from the legitimate user. Therefore, it can be used to find the best threshold for classifying the current user into an impostor or legal user. SVR will be used to calculate the user's input score from zero to one. Zero indicates an impostor while one identifies the legal user. We set the optimum threshold between 0 and 1 to classify the user's input as that of the lowest EER. If the user's input score is below the threshold, the result indicates an impostor. Otherwise, it will be the legal user. Additionally, we choose SVR not only for its great accuracy and robustness but also for its ease of use and quick calculation.

B. ROBUSTNESS ENHANCEMENT

The authentication model will be passed on to the robust testing process. Our experiment shows that different physical activities are associated with different but essentially similar biometric patterns such that biometric models may be disturbed by body motion conditions. We construct adversarial samples according to the distribution of each feature in the original samples. In the robustness test, the adversarial noise will only be added to the validation samples to test whether the EER will decrease or not. If the EER goes down, this indicates that the model is not sufficiently robust for all body motion conditions and we will retrain the model with AN samples.

C. EVALUATION METHOD

Considering the practicability of authentication, we want to build models with different levels of knowledge of impostors. We have three levels: knowing all impostors, knowing half of the impostors, and knowing none of the impostors. We apply the verification process (VP) introduced in Mondal and Bours's [37] work. In verification process 1 (VP1), all impostors are included in the training process as shown in Figure 2. We then calculate the result using validation samples from known users and impostors. In verification process 1 (VP1), all the impostors' typing behavior has been learned in the training process as shown in Figure 2. In verification process 2 (VP2), half of the impostors' typing behavior has been learned in the training process as shown in Figure 3. In verification process 3 (VP3), none of the impostors' samples are included in the training set as shown in Figure 4. VP3 is more in line with the real-life scenario. All VPs have the same number of authorized users.

target user train	target user validation	target user test
No.1 impostor train	No.1 validation	No.1 test
No.2 impostor train	No.2 validation	No.2 test
No.? impostor train	... validation	... test
No.m impostor train	No.m validation	No.m test

FIGURE 2. Verification process 1.

target user train	target user validation	target user test
No.1 impostor train	No.1 validation	No.1 test
No.? impostor train	... validation	... test
No.m/2 impostor train	No.m/2 validation	No.m/2 test
		No.(m/2)+1 test
		... test
		No.m test

FIGURE 3. Verification process 2.

target user train	target user validation	target user test
No.1 impostor train	No.1 validation	
No.? impostor train	... validation	
No.m/2 impostor train	No.m/2 validation	
		No.(m/2)+1 test
		... test
		No.m test

FIGURE 4. Verification process 3.

We use the Equal Error Rate (EER) to evaluate the performance of our model. EER is an algorithm used to predetermine the threshold for the rate of false acceptances and false rejections. The False Acceptance Rate (FAR) is the propensity of a security system to mistakenly verify an unauthorized person while the False Rejection Rate (FRR) is the propensity of a security system failing to admit an authorized person. The EER is the point at which the FAR and FRR are equal. They are defined as follows:

$$FAR = \frac{Num(\widehat{y} == 1 \&\& y == 0)}{Num(y == 0)} \tag{1}$$

$$FRR = \frac{Num(\widehat{y} == 0 \&\& y == 1)}{Num(y == 1)} \tag{2}$$

where Num stands for the number of cases. $\widehat{y}$ represents the predicted value and y is the actual value.

IV. DATASET AND METHODS

We enlisted 104 participants for our experiments and used PIN codes 1-1-1-1, 3-2-4-4, and 5-5-5-5. Most of the participants were college students aged between 20 and 25 as shown in Table 1. The number of male and female participants was 47 and 57 respectively. Participants in our study were asked to enter an error-free PIN code at least 20 times, which means some of them may input more than 20 attempts if they wished. We collected a total of 6311 error-free actions. Our application automatically records the user's acceleration,

TABLE 1. Participants information.

Age group	Female	Male	Total
21-25	43	53	96
26-30	2	2	4
31-35	2	0	2
36-40	0	1	1
41-45	0	1	1
Total	47	57	104

pressure, size, and time data during the process. We refer to an action as a complete instance of a password input.

A. FEATURES

We build a more comprehensive feature set through the current feature set research. Based on the raw data, we arranged the features into four groups: time related, acceleration, pressure, and size. We extracted more time-oriented features as Sheng *et al.* [38] did in their work, as shown in Figure 5. We also believe that all the acceleration measurements have the potential to provide useful information, not just their minimum, maximum and mean values. Therefore, we refined the feature groups by adding three more groups of acceleration values in each spatial axis. As a result, the number of features increased from 63 to 146. We describe them as follows:

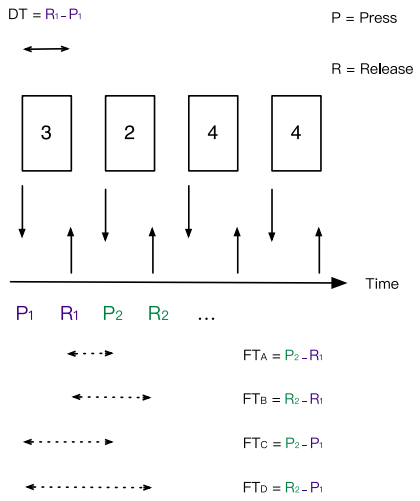


FIGURE 5. Time feature.

- **Time**
  We describe time features in Figure 5. We divide one 4-digit PIN code into two parts. For "3-2-4-4", the first two digits "3-2" contain time features including 'dwell time' (DT), 'flight time' FT A, FT B, FT C and FT D as shown in Figure 5. In total, there are 10 time features in each trial.
- **Acceleration**
  We record acceleration features in two parts. The first part consists of acceleration in each axis for every digit. Most of the key-hold times can have more than three groups of accelerations sampled at 100Hz. We record the first three groups of acceleration features during the key-hold time. As we have four digits, three axes, and three groups of linear and angular accelerations, we have 72 features in the first part of the PIN. The second part contains the press time, release time, maximum, minimum, average, and standard deviation of accelerations for every digit. Therefore, we get 48 features for the second part, and 120 acceleration features are recorded in total.
- **Pressure**
  There are four pressing and four releasing times in a 4-digit PIN code. So, we record eight pressure features in each trial.
- **Size**
  Each time the user presses or releases a key, the mobile device receives a press size feature. So, there are eight size features in each trial.

B. FEATURE SELECTION

In order to optimize our extensive feature set, we want to find the best feature selection method. We choose the most typical methods and conduct experiments as shown in Table 2 and Table 3. For a given measure of quality, DE can optimize solutions to a problem by attempting to improve solutions iteratively. It holds a population of candidate solutions and creates new candidates by combining existing ones according to an algorithm. Then, it retains the solution which gives the best results on the optimization problem. We selected DE for the following three reasons: 1) Alternative feature ranking methods like Pearson [24] and Chi-square [39] only consider the relationship between the feature and the result, but DE also takes the relations between features into account; 2) genetic algorithms (GA) [40] and Binary Particle Swarm Optimization (BPSO) [26] can also be used to select the best group of features, but the influence of population size on processing time in GA is exponential; 3) the solution time for BPSO and DE increases linearly with population size [41]

TABLE 2. Feature selection comparison.

Group	Input	Baseline EER	SVR EER
The original 63 features	1111	23.3333	8.33333
	3244	12.0895	3.06371
	5555	21.3333	7.33333
The new 146 features	1111	11.64998	4.98614
	3244	8.49031	2.71304
	5555	9.62441	4.88135
Pearson feature selection (top 100)	1111	15.8333	5.27778
	3244	9.65802	2.71304
	5555	10.38889	5.16667
BPSO feature selection	1111	11.6667	3.10249
	3244	6.19114	0.43582
	5555	15.1807	1.66667
DE feature selection	1111	7.5	**2.11911**
	3244	6.66667	**0.12660**
	5555	9.84241	**0.77692**

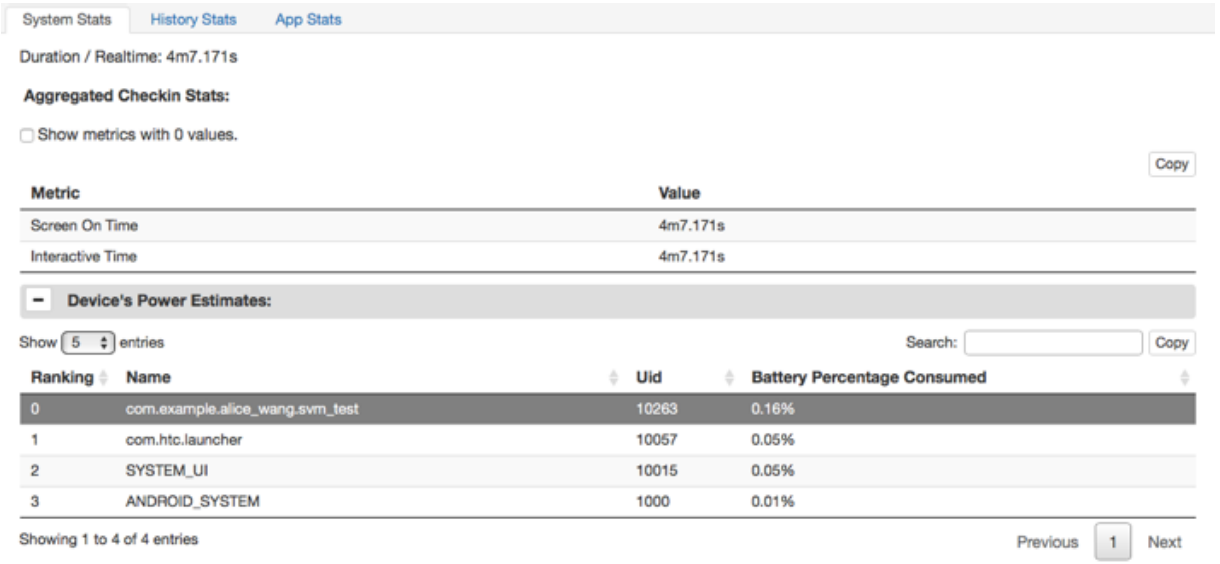


FIGURE 6. Battery consumption of SVR method with 146 features.

TABLE 3. Time test on BPSO and DE.

Group	Number of particles(agents)	Number of iterations	Time
BPSO	100	10	2,273s
DE	100	10	1,919s

and out of these two, the reason why we do not choose BPSO is that BPSO is slower than DE and it does not perform well as DE.

In our DE method, each chromosome has a list of 146 random integers valued 0 or 1, in which 0 stands for discarding this feature and 1 means applying this feature in later classification work. Our goal is to find the most significant feature group that gives the lowest EER on the validation samples. We experimented to verify whether the combination of more features with DE feature selection is a good choice. As shown in Table 2, the original 63 features can only obtain an EER of 3.06371% for 3-2-4-4, but 146 features can reach an EER of 2.71304%. Then, after using the feature selection methods, further DE selection can provide the best EER of 0.12660% compared with Pearson and BPSO. In Table 2, we highlight the best EER for each PIN code with bold font and underline.

The results of the comparison between different feature selection methods are shown in Table 2. We select the DE feature selection method on this basis. DE can also help to reduce the energy consumption. System files record the battery consumption rate and tools like Battery Historian [42] can use this data to analyze battery usage. We record it as the accumulation of battery consumption after 20 trials on the Android system. Battery Historian can help us analyze the system file and determine the battery consumption as shown in Figure 6. We tested the phone with 146 features and found the battery consumption to be 0.16% according to the system file. After the DE method was applied, this reduced to 0.11%. Hence, the battery consumption could be reduced by 31.25% with the DE method because fewer features are utilized.

C. ADVERSARIAL NOISE

It is hard and inconvenient to collect data for all body motions, but we can analyze keystroke data distribution in different contexts and simulate the original samples with adversarial noise(AN) samples. To deal with authentication for different body motion conditions, we provide one model for all activity types because activity recognition may introduce errors and use a lot of energy if we record the phone status continuously. We had the same user enter 3-2-4-4, 20 times in four situations: lying, sitting, standing, and walking. We then selected ten random features in different body motion conditions as shown in Figure 7. We can see from the figure that though different body motion conditions have different feature distributions, they are alike in some key regards. Therefore, we apply adversarial noise according to these factors to represent the user's individual pattern.

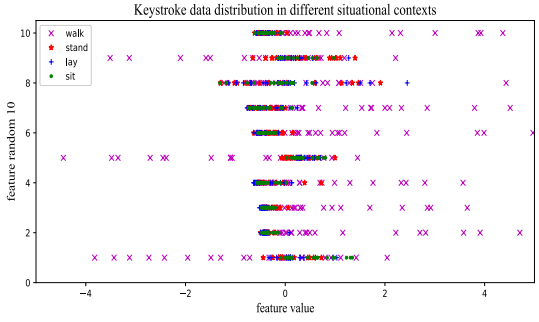


FIGURE 7. Keystroke data distribution in different situational contexts.

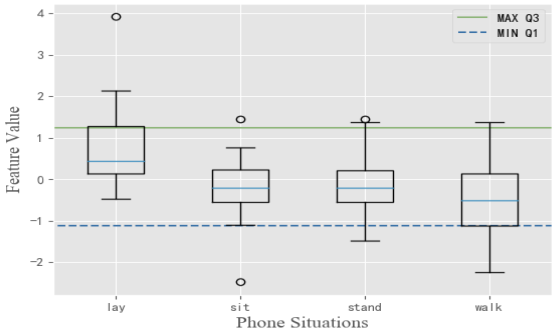


FIGURE 8. Data Noise Boxplot.

We randomly selected one feature for analysis in Figure 8. Q1 represents the 25th percentile of the feature's statistical distribution. Q3 represents the 75th percentile of the feature's distribution. The interquartile range (IQR), also known as the mid-spread or middle 50%, is equal to the difference between the 25th and 75th percentiles, IQR = Q3 - Q1. Since the IQR for each feature in each body motion condition can represent the user's behavioral pattern, the union of IQRs for each feature in all body motion conditions is also a part of the user's behavioral pattern. Adversarial noise is defined as a random value between the minimum of Q1 in 4 situations and the maximum of Q3 in 4 situations, as shown in the equation:

$$AN_j = Random(\bigcup_{n=lay,sit,stand,walk} (Q1_n, Q3_n)) \quad (3)$$

where j belongs to the group of feature IDs after DE selection. As IQR is a trimmed estimator and is the most significant robust measure of scale, we believe the union of IQRs for each feature in all body motion conditions is also robust for the user's keystroke biometrics.

V. AUTHENTICATION EXPERIMENTS

Our authentication process is shown in Figure 9. We first evaluate the performance of the authentication model, and record the result as *EER*. Then validation samples with AN are also tested with the authentication model and the result is recorded as *EERwithAN*. If *EERwithAN* is larger than *EER*, the model will be retrained with AN to improve its robustness. If *EERwithAN* is smaller than *EERwithAN_Pre*, the scheme will continue to retrain the model. The loop will end if *EERwithAN* is larger than *EERwithAN_Pre*. After its robustness has been verified, the authentication model will then be passed on to the test process. We will describe these procedures in more detail in the following subsections with a default input of 3-2-4-4.

A. AUTHENTICATION ACCURACY

Table 4 shows the EERs from Zheng *et al.*'s [23] work and the results from our data using Zheng's method as a baseline along with results using other classification methods, including SVR, Scaled Euclidean(SE) [43], Scaled Manhattan(SM) [43], KNN [33], and RF [32]. These classifiers are used in Sitová *et al.*'s [20] work and Kambourakis *et al.*'s [31] work. The kernel used in the SVR model is "rbf", we use a coefficient of 0.0 and a gamma of 0.00685. In Table 4, "(a)" indicates using all 146 features and "(b)" represents using the selected features after DE. We highlight the lowest EER for every row in bold font and the lowest EER for each PIN code with an underline.

In VP1, we assume that all impostors are known to the model. After the DE phase, the SVR method gives the best EER of 0.12660% which is 98.10% lower than the EER of 6.66667% in the baseline method. The baseline method has the advantage in the VP2 and VP3 stages that it chooses thresholds using the distance from each sample to the nearest legal user's sample. As it largely relies on the target user's training samples, if we can find the maximum distance of every legitimate user's sample to its nearest legitimate sample, not including itself, we can directly choose the maximum distance to be a good threshold. However, the baseline method was found to have poor generalization ability for different body motion conditions in the robustness test.

Table 4 shows the value of every EER for each PIN code. Figure 10 describes the trade-off between FRR and FAR in all three PIN combinations using SVR. There are six figures altogether, showing false rates and thresholds for every PIN code in each verification process before and after the DE phase. We deem an input action with a score above the threshold to come from a legal user. Otherwise, it is from an impostor.

B. ROBUSTNESS TEST

To handle different input positions, we analyze multiple samples from different body motion conditions and create adversarial noise samples according to the original samples. Then, the adversarial noise samples are added to the validation set. We can assess whether or not the model is robust enough to handle all body motion conditions through the new validation set.

Table 5 shows the robustness in VP3 for models both with and without adversarial noise test samples. The lowest EER for each PIN code in every row is emphasized with bold font, and we highlight the best EER, with AN test samples, using underscores. Here, the EER has increased by 173.45% from the baseline. However, the EER with SVR has only increased by 83.59%, so SVR has a better generalization ability for all body motion conditions. Figure 11 shows the false rates and thresholds with adversarial noise validation samples.

C. STRENGTHENING ROBUSTNESS

As shown in the previous experiment, body motion conditions can have a significant influence on the robustness of the model. Hence, AN samples are added to

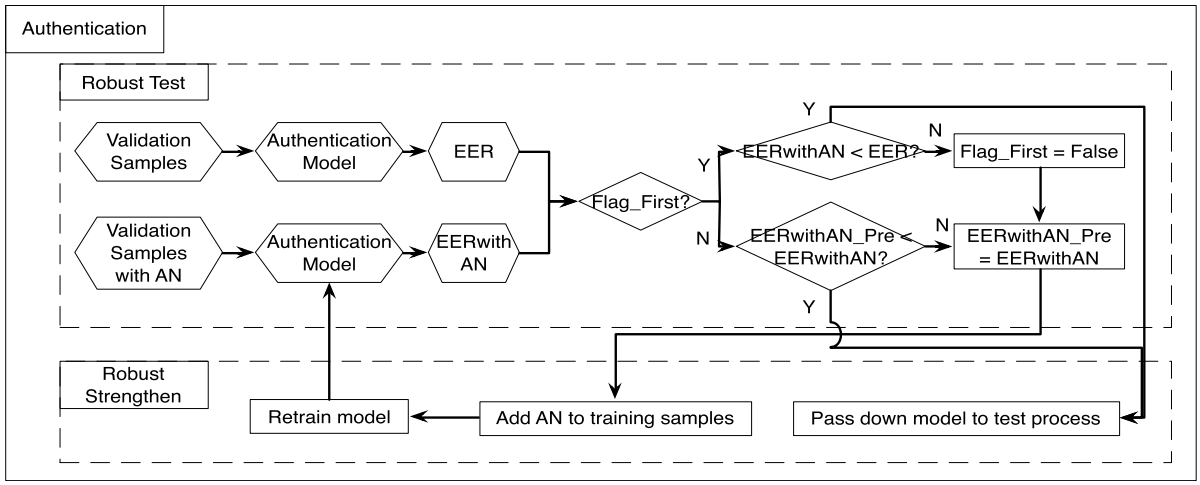


FIGURE 9. Authentication Flow Chart.

TABLE 4. EER comparison.

Group	Input	Baseline EER	SVR EER	SE EER	SM EER	KNN EER	RF EER
Zheng et al.	1111	6.96					
	3244	3.65					
	5555	7.34					
VP1(a)	1111	11.6499	**4.98615**	24.6665	15.0	28.3333	23.3333
	3244	8.49030	**2.71305**	14.9999	12.5	16.25	18.75
	5555	9.62441	**4.88136**	27.3332	20.0	21.6667	18.3333
VP1(b)	1111	7.5	**2.11911**	23.3332	15.0	26.25	21.6667
	3244	6.66667	**0.12660**	11.5512	11.25	15.0	10.0
	5555	9.84242	**0.77692**	27.5712	20.3689	23.3333	16.6667
VP2(a)	1111	**11.6499**	12.0086	28.3949	14.9998	20.3689	21.9884
	3244	**8.49030**	8.98855	14.9999	12.4999	14.2777	18.9501
	5555	**9.62441**	11.7491	29.9997	19.9998	16.4999	19.2417
VP2(b)	1111	**7.5**	7.76408	27.9997	14.0454	21.9021	24.3059
	3244	**6.66667**	6.83333	12.4999	11.0645	12.2774	18.1958
	5555	**9.84242**	10.4389	27.9997	21.7064	10.9444	20.7154
VP3(a)	1111	**12.2082**	13.3333	27.9939	15.2714	24.6728	25.5853
	3244	9.18703	**8.75**	14.9999	12.4999	16.6209	23.3569
	5555	10.5772	**10.0**	29.5793	19.9998	16.4999	21.9813
VP3(b)	1111	7.97208	**6.46302**	23.7816	17.1230	26.4843	29.9687
	3244	6.66667	**6.35464**	13.8804	12.4999	14.8869	22.3433
	5555	10.2050	**10.0**	28.3299	24.9997	10.9444	23.6110

VP = Verification Process, a = 146 features, b = feature sequence by DE

the training samples to help the model perform better in different circumstances. Each group of AN samples is the same size as the original training set. We add AN samples group by group to find the best authentication model with the lowest EER as shown in Table 6. When the number of AN sample groups reaches three, the EER stops declining.

In this table, "(a)" represents the original training samples, and "(b-e)" stands for the combination of "(a)" and AN samples. "*2" indicates the number of AN the sample group. "(d)" consists of the original training samples and three groups of adversarial samples have the best EER of 8.70932% for 3-2-4-4, 9.42637% for 5-5-5-5 and 10.0965% for 1-1-1-1.

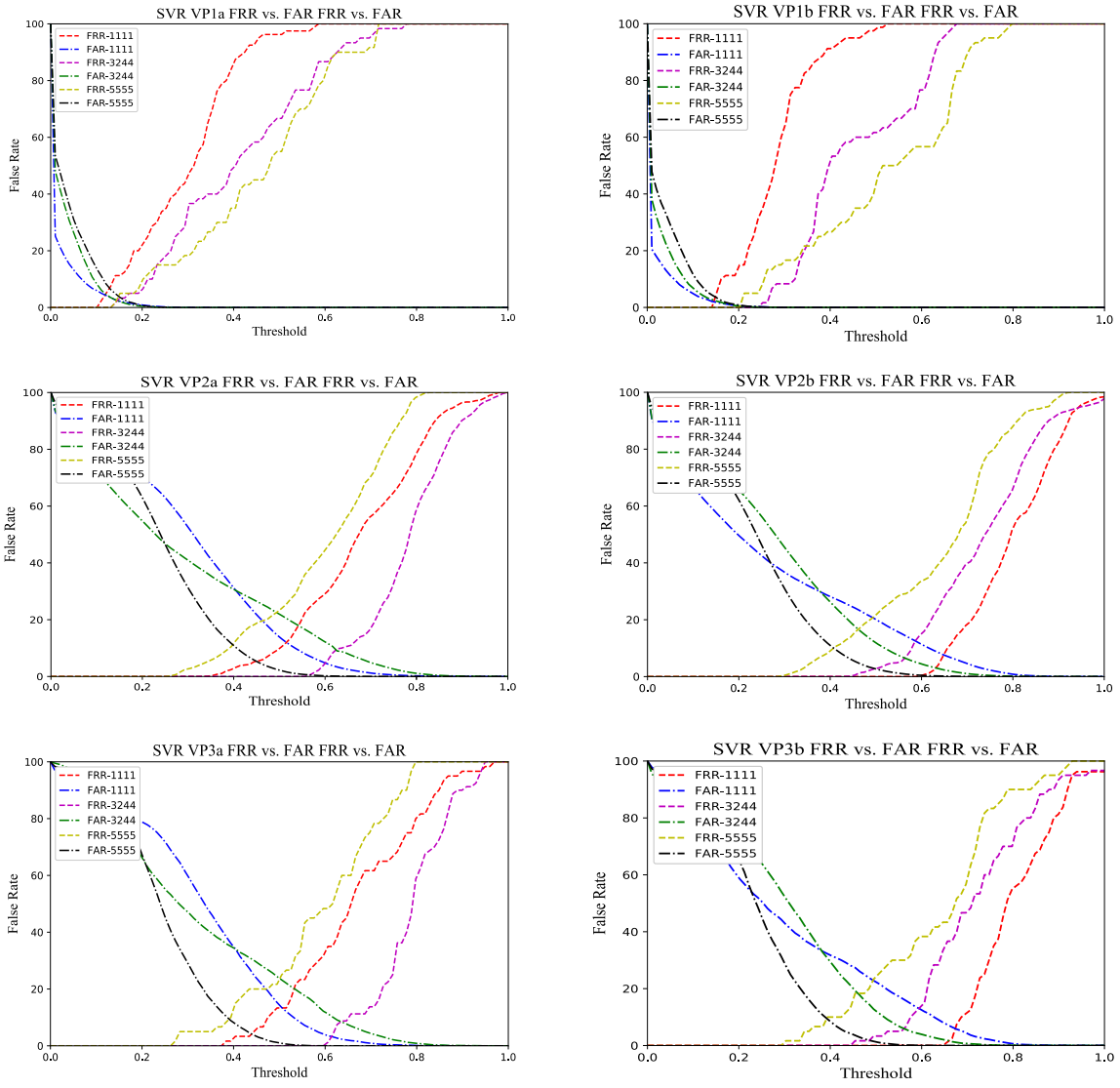
TABLE 5. EERs with and without AN (validation only).

Group VP3b	Input	Baseline	SVR
EER without perturbations	1111	7.97208	**6.46302**
	3244	6.66667	**6.35464**
	5555	10.2050	**10.0**
EER with adversarial noise only in test samples	1111	21.3888	**10.8601**
	3244	18.2302	**10.9299**
	5555	24.4282	**11.6667**

VP = Verification Process, a = 146 features, b = feature sequence by DE

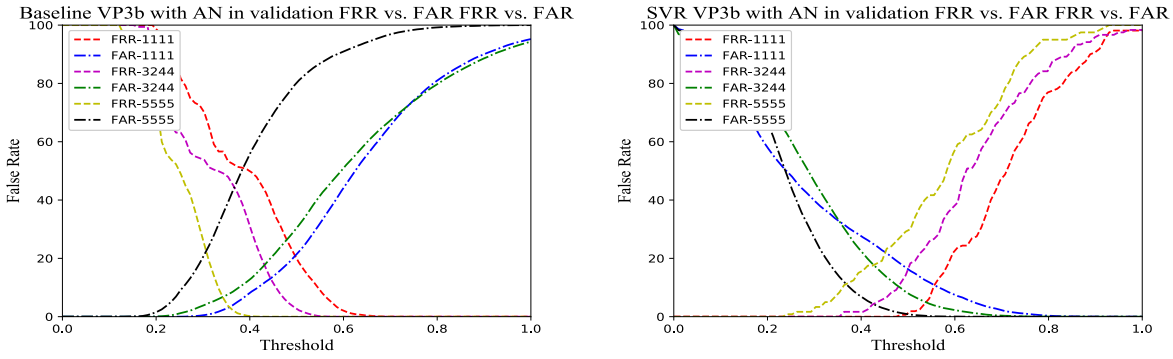
The false rates and thresholds of "(d)" and "(e)" with the SVR method are shown in Figure 12.

It can be seen from Table 6 and Figure 12 that almost all the models have a slight improvement after retraining



VP = Verification Process, a = 146 features, b = feature group selected by DE

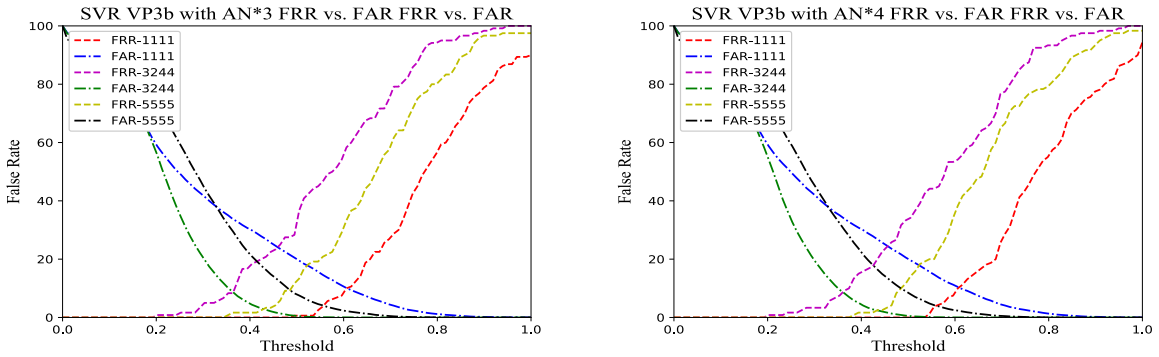
FIGURE 10. False Rates and Thresholds in SVR. VP = Verification Process, a = 146 features, b = feature group selected by DE.



VP = Verification Process, a = 146 features, b = feature group by DE

FIGURE 11. EERs with AN (validation only). VP = Verification Process, a = 146 features, b = feature group by DE.

with samples of various body motion conditions, not only the SVR model. Also, SVR has the best performance in EER for the PIN 1-1-1-1, a reduction of 7.03%, and the EER for 5-5-5-5 decreases by 19.20%. The most surprising result is the EER of 8.70932% for 3-2-4-4, which is reduced by 20.32%.



VP = Verification Process, a =146 features, b =feature group by DE, AN*n = train samples with n groups of AN

FIGURE 12. EERs with AN in SVR. VP = Verification Process, a = 146 features, b = feature group by DE, AN*n = train samples with n groups of AN.

TABLE 6. EERs with AN.

Group VP3b with AN test samples	Input	Baseline	SVR
(a) Train samples	1111	21.3888	**10.8601**
	3244	18.2302	**10.9299**
	5555	23.8474	**11.6667**
(b) a+AN train samples*1	1111	20.625	**10.3054**
	3244	18.0550	**10.1892**
	5555	23.3333	**10.7582**
(c) a+AN train samples*2	1111	20.5466	**10.5627**
	3244	17.7170	**9.16667**
	5555	23.1047	**10.0927**
(d) a+AN train samples*3	1111	19.6061	**10.0965**
	3244	17.7090	**8.70932**
	5555	22.8939	**9.42637**
(e) a+AN train samples*4	1111	20.7637	**10.4662**
	3244	17.6205	**9.16667**
	5555	23.3333	**10.0**

VP = Verification Process, a = 146 features, b = feature group by DE,AN train samples*n = train samples with n groups of AN

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the DEANUA scheme, which focuses on accuracy and robustness in smartphone keystroke dynamics. We conduct experiments to obtain 6311 4-digit PIN code trials. During the feature selection process, a DE technique is applied to select useful features from all 146 candidates. The best EER obtained from the SVR was 0.12660% for the code 3-2-4-4 in VP1, which provides a 98.10% reduction in ERR compared with the baseline method. The DE feature selection method also reduces energy consumption by 31.25%. In many applications, minimizing FAR is more important than FRR. When the lowest FAR of 0.02817% is obtained, the FRR reaches 1.66667%. While this still does not meet the EU recommendations, we have demonstrated progress toward it in our mobile biometric authentication study.

The DEANUA scheme has not only improved the accuracy but also enhances robustness in all body motion conditions. Adversarial noise, which is created according to the distribution of samples in different contexts can lead to misclassifications. Hence, we added adversarial noise to the training samples to help the model perform better in different body motion conditions. Finally, the DEANUA scheme achieved an EER of 8.70932% in VP3, which means we can protect the user without prior knowledge of the impostors, in all body motion conditions.

As for future work, data can be recorded not only under laboratory conditions but also in the real world, and user identification with keystroke dynamics can satisfy the need for recognizing multiple users of the same mobile device.

REFERENCES

[1] S. Khan, M. Nauman, A. T. Othman, and S. Musa, "How secure is your smartphone: An analysis of smartphone security mechanisms," in *Proc. IEEE Int. Conf. Cyber Secur., Cyber Warfare Digit. Forensic*, Jun. 2012, pp. 76–81.

[2] W. Jeon, J. Kim, Y. Lee, and D. Won, "A practical analysis of smartphone security," in *Proc. Symp. Hum. Interface*, 2011, pp. 311–320.

[3] IDC2017. (2017). *Smartphone OS Market Share*. [Online]. Available: https://www.idc.com/promo/smartphone-market-share/os

[4] B. Arnaud and N. L. Clarke, "Deployment of keystroke analysis on a smartphone," in *Proc. Austral. Inf. Secur. Manage. Conf.*, 2008, p. 48.

[5] T. Feng, J. Yang, Z. Yan, E. M. Tapia, and W. Shi, "TIPS: Context-aware implicit user identification using touch screen in uncontrolled environments," in *Proc. ACM 15th Workshop Mobile Comput. Syst. Appl.*, 2014, p. 9.

[6] A. Toosi, A. Bottino, S. Cumani, P. Negri, and P. L. Sottile, "Feature fusion for fingerprint liveness detection: A comparative study," *IEEE Access*, vol. 5, pp. 23695–23709, 2017.

[7] S. Chinchu, A. Mohammed, and B. S. Mahesh, "A novel method for real time face spoof recognition for single and multiple user authentication," in *Proc. IEEE Int. Conf. Intell. Comput., Instrum. Control Technol.*, Jul. 2017, pp. 376–380.

[8] DataGenetics. (2017). *Pin Analysis*. [Online]. Available: http://www.datagenetics.com/blog/september32012/

[9] Y. Xu, J. Heinly, A. M. White, F. Monrose, and J. M. Frahm, "Seeing double: Reconstructing obscured typed input from repeated compromising reflections," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2013, pp. 1063–1074.

[10] C. Sousedik and C. Busch, "Presentation attack detection methods for fingerprint recognition systems: A survey," *IET Biometrics*, vol. 3, no. 4, pp. 219–233, 2014.

[11] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2013, pp. 1–6.

[12] A. Alzubaidi and J. Kalita, "Authentication of smartphone users using behavioral biometrics," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1998–2026, 3rd Quart., 2016.
[13] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Generat. Comput. Syst.*, vol. 16, no. 4, pp. 351–359, 2000.
[14] M. Shahzad, A. X. Liu, and A. Samuel, "Behavior based human authentication on touch screen devices using gestures and signatures," *IEEE Trans. Mobile Comput.*, vol. 16, no. 10, pp. 2726–2741, Oct. 2017.
[15] A. Buriro, B. Crispo, F. Del Frari, and K. Wrona, "Touchstroke: Smartphone user authentication based on touch-typing biometrics," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2015.
[16] P. S. Teh *et al.*, "TDAS: A touch dynamics based multi-factor authentication solution for mobile devices," *Int. J. Pervasive Comput. Commun.*, vol. 12, no. 1, pp. 127–153, 2016.
[17] A. Buriro, Z. Akhtar, B. Crispo, and F. Del Frari, "Age, gender and operating-hand estimation on smart mobile devices," in *Proc. Int. Conf. IEEE Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.
[18] H. Saevanee and P. Bhatarakosol, "User authentication using combination of behavioral biometrics over the touchpad acting like touch screen of mobile device," in *Proc. IEEE Int. Conf. Comput. Elect. Eng.*, Dec. 2008, pp. 82–86.
[19] Y. Meng and D. S. Wong, "Design of touch dynamics based user authentication with an adaptive mechanism on mobile phones," in *Proc. ACM Symp. Appl. Comput.*, 2014, pp. 1680–1687.
[20] Z. Sitová *et al.*, "HMOG: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 877–892, May 2016.
[21] H. Crawford and E. Ahmadzadeh, "Authentication on the go: Assessing the effect of movement on mobile device keystroke dynamics," in *Proc. USENIX Assoc.*, 2017, pp. 163–173.
[22] *Alarm systems. Access control systems for use in security applications. Part 1: System requirements*, European Standard EN 50133-1, Standard Number EN 50133-1:1996/A1:2002, Technical Body CLC/TC 79, European Committee for Electrotechnical Standardization (CENELEC), 2002.
[23] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on smartphones via tapping behaviors," in *Proc. IEEE 22nd Int. Conf. Netw. Protocols*, Oct. 2014, pp. 221–232.
[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[25] R. Storn and K. Price, "Differential evolution–A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[26] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. Comput. Cybern. Simulation*, Oct. 1997, pp. 4104–4108.
[27] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[28] E. Maiorana, P. Campisi, N. González-Carballo, and A. Neri, "Keystroke dynamics authentication for mobile phones," in *Proc. ACM Symp. Appl. Comput.*, 2011, pp. 21–26.
[29] M. Trojahn and F. Ortmeier, "Toward mobile authentication with keystroke dynamics on mobile phones and tablets," in *Proc. IEEE Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2013, pp. 697–702.
[30] C. Giuffrida, K. Majdanik, M. Conti, and H. Bos, "I sensed it was you: Authenticating mobile users with sensor-enhanced keystroke dynamics," in *Proc. Int. Conf. Detect. Intrusions Malware, Vulnerability Assessment*, 2014, pp. 92–111.
[31] G. Kambourakis, D. Damopoulos, D. Papamartzivanos, and E. Pavlidakis, "Introducing touchstroke: Keystroke-based authentication system for smartphones," *Secur. Commun. Netw.*, vol. 9, no. 6, pp. 542–554, 2016.
[32] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2007.
[33] T. Denoeux, "A K-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst., Man and*, vol. 25, no. 5, pp. 804–813, May 1995.
[34] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.
[35] A. Morales *et al.*, "Keystroke biometrics ongoing competition," *IEEE Access*, vol. 4, pp. 7736–7746, 2016.
[36] M. Ataş, "Hand tremor based biometric recognition using leap motion device," *IEEE Access*, vol. 5, pp. 23320–23326, 2017.
[37] S. Mondal and P. Bours, "A computational approach to the continuous authentication biometric system," *Inf. Sci.*, vol. 304, pp. 28–53, May 2015.
[38] Y. Sheng, V. V. Phoha, and S. M. Rovnyak, "A parallel decision tree-based method for user authentication based on keystroke patterns," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 4, pp. 826–833, Aug. 2005.
[39] H. Scheffé, "The relation of control charts to analysis of variance and chi-square tests," *J. Amer. Statist. Assoc.*, vol. 42, no. 239, pp. 425–431, 1947.
[40] J. H. Holland and D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Massachusetts, MA, USA: Addison-Wesley, 1989.
[41] V. Kachitvichyanukul, "Comparison of three evolutionary algorithms: GA, PSO, and DE," *Ind. Eng. Manage. Syst.*, vol. 11, no. 3, pp. 215–223, 2012.
[42] *Google Battery-Historian*. [Online]. Available: https://github.com/google/battery-historian
[43] K. S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *Proc. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun./Jul. 2009, pp. 125–134.

YUHUA WANG received the B.Eng. degree in computer science and engineering from the University of Electronic Science and Technology of China, Sichuan, China, in 2016. She is currently pursuing the master's degree in information security with the Beijing University of Posts and Telecommunications. She has participated in one of the National Key R&D Programs of China and one of the Science Foundations of China. Her research interests include the information security, biometrics, and pattern recognition techniques.

CHUNHUA WU received the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, where she is currently an Assistant Professor with the School of Cyberspace Security. She has authored articles in many journals and conferences. She is in charge of one of the National Natural Science Foundation of China. Her current research interests include machine learning methods to solve problems in network security, including attack detection, malware detection, and identity authentication.

KANGFENG ZHENG received the Ph.D. degree in information and signal processing from the Beijing University of Posts and Telecommunications, in 2006, where he is currently an Associate Professor with the School of Cyberspace Security. His research interests include networking and system security, network information processing, and network coding.

XIUJUAN WANG received the Ph.D. degree in information and signal processing from the Beijing University of Posts and Telecommunications, in 2006. She is currently an Instructor Lecturer with the Faculty of Information Technology, Beijing University of Technology. Her research interests include information and signal processing, network security, and network coding.