2IN1: A Bimodal Behavioral Biometric-based User Authentication Scheme for Smartphones

Attaullah Buriro\textsuperscript{1}, Flaminia Luccio\textsuperscript{1}

\textsuperscript{1}Department of Environmental Sciences, Informatics and Statistics, Ca’ Foscari University of Venice, Via Torino, 155, Venice, Italy

Abstract
This paper introduces a bi-modal mechanism that leverages the way a smartphone user signs on the touchscreen and taps/enters any “text-independent” 8-digit numbers to authenticate their identity. Precisely, by extracting the trajectory of touch-points and touch-timing features during the enrollment stage, our scheme creates a digital identity of a user based on these behaviors. In the verification stage, our scheme compares the captured touch-points and touch-timing signatures with the digital identity of the user created during the enrollment stage. If the captured signatures match the digital identity within a certain tolerance, the user is authenticated. The choice of low-level events, such as signing on the screen and touch-typing, as biometric modalities makes our scheme easier to implement and adapt. We evaluated our approach using multiple classifiers, i.e., K-Nearest Neighbor, Support Vector Machine, and Deep Neural Network, and achieved a high True Acceptance Rate of 97.1\% with a low False Acceptance Rate of just 0.2\%, and an accuracy of 98.45\% on a dataset of 20 volunteers. These results prove our scheme accurate in verifying the identity of users while also maintaining a low rate of false acceptance of unauthorized users.

Keywords
Behavioral Biometrics, Smartphone User Authentication, Keystroke Dynamics, Signature Recognition

1. Introduction

Smartphones are personal portable devices mainly designed for communication, such as making phone calls and sending audiovisual messages. However, because they offer anytime-anywhere connectivity and computing capabilities, they are used to perform a range of sensitive tasks. i.e., social networking, mobile/internet banking, entertainment, fitness tracking, etc. In general, smartphones are designed to be an all-in-one solution for communication and computing needs. Hence, access to the smartphone and the data it contains must be limited to its owner/legitimate user only.

Smartphone user authentication is the process of verifying the identity of someone attempting to access it. Existing authentication methods based on PINs and passwords are not convenient for the type of user interactions that characterize smartphones (very frequent \cite{1} and short \cite{2}), resulting in more and more users not using any security (up to 40.9\% according to a recent study \cite{3}). As a result, the focus of security research has shifted towards biometric-based au-
thentication schemes, with behavioral biometrics being particularly attractive because it is easy to implement and only requires standard hardware provided by most modern smartphones [4].

Multi-modal behavioral biometric solutions, compared to uni-modal systems, are considered better because they have shown to be highly accurate and secure [5]. A multi-modal system leverages multiple methods, such as face, fingerprint, etc., to confirm a user’s identity. This combination makes extremely difficult for an attacker to bypass the system, as they need to mimic multiple forms of authentication. Additionally, using multiple methods of identification can also increase the overall accuracy of the system, as some users may have difficulty with one form of identification but not another [6].

This paper presents a bi-modal smartphone user authentication scheme, which leverages two inherently secure and unique human behaviors: (i) touch-typing¹, i.e., the way a user taps a combination of 8-digits “text-independent” number, not necessarily a secret, and (ii) signature². Touch-typing and signature are often used as behavioral modalities in multi-modal biometric systems because users are familiar with them and they are easy to use. For example, a signature is a well-adapted behavioral action in our daily lives, and many users are accustomed to it. Touch-typing, or typing on a keyboard, is also a familiar task for many people and can be used to confirm a user’s identity by analyzing their typing rhythm, speed, and pressure. Our scheme, by leveraging these two behaviors, becomes a more secure and reliable way to verify a user’s identity.

We evaluated our scheme on our collected dataset of 20 users by applying a multiclass classification approach. A multiclass classification approach involves training a model to distinguish between three or more classes of instances, in this case, users. In this case, the goal is to identify the user’s identity based on their behavior during authentication. More technically, our scheme aims to address the challenge of user authentication within client-server architectures, such as those used in banking and remote access systems. We chose multiple classical classifiers, such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), as baseline models, including a Deep Neural Network (DNN) as multiclass classifiers to evaluate our scheme. Results show that the DNN classifier outperformed other chosen classifiers and performed well, attaining a higher accuracy of 98.45% and a lower False Acceptance Rate (FAR) of just 0.2%.

The main contributions of this work are:

- The proposal of a bi-modal smartphone user authentication scheme, which authenticates the user based on the differences in tap-timings (while she taps/enters 8-digit “text-independent” number), and the signature features.
- The adaptation of DNN as a classifier.
- Proof-of-the-concept implementation of the proposed solution on a real Android smartphone.

Paper Organization: Section 2 surveys the most relevant papers in the field and discusses the motivations for improving user authentication. Section 3 explains the proposed scheme.

¹Touch-typing refers to the way a user types on their smartphone’s virtual keyboard. This can include things like the speed and accuracy of the typing, as well as the pressure applied to the keys.
²Signature as a behavior refers to the way a user writes on a smartphone touchscreen. The features of interest could be the velocity of writing, pressure being applied, and the covered Euclidean distance.
Section 4 describes the detailed methodology used in our experiments, including protocols for data collection, feature extraction and selection, and analysis. Section 5 explains our obtained results. Section 6 summarizes the findings of the paper and discusses future work.

2. Related Work

In this section, we survey the most relevant published papers.

2.1. Unimodal Systems

2.1.1. Touch-typing User Authentication

Touch-typing biometrics is a method of user authentication that exploits the unique patterns of a person’s typing rhythm based on unique features such as the speed, pressure, and duration of each keystroke and style to verify their identity. Though touch-typing has emerged as a secure and usable biometric recently [7] [8], it is still relatively a new and emerging technology, hence it still requires more research and development to be widely used.

Kambourakis et al. [8] explore the potential of fixed text keystroke dynamics for user authentication on touchscreen-equipped smartphones. They discovered two novel features, i.e., speed and distance, and demonstrated their contribution to obtaining accuracy. In the best-case scenario, using KNN as a classifier, authors reported a FAR of 23.7% and False Reject Rate (FRR) of 3.5%. The paper by Zheng et al [9] proposes a user verification system for smartphones that uses the tapping behaviors of the users as a behavioral biometric. The system extracts four features from the smartphone sensors: acceleration, pressure, size, and time. The system then uses a support vector machine (SVM) classifier to determine whether the user who enters the passcode is the true owner of the smartphone or an impostor. The paper evaluates the system on a dataset of over 80 users and reports an average Equal Error Rate (EER) of 3.65%.

2.1.2. Signature-based User Authentication

Signature-based user authentication schemes [10][11] for smartphones are methods that use the handwritten signature of the user as a biometric modality to verify their identity. These schemes aim to provide a natural and convenient way of authentication, as well as a high level of security. Signature-based authentication schemes can be classified into two categories: static and dynamic. Static schemes capture the signature image as a whole and compare it with a stored template, while dynamic schemes capture the temporal and spatial features of the signature, such as speed, pressure, and acceleration, and use them for verification.

Yang and Liu [10] focus on online signature verification using wavelet packet analysis to extract dynamic local features while integrating global features to preserve the integrity of signature data. To address limitations associated with traditional expectation maximization (EM) algorithms, which are prone to dependency on parameter initialization and susceptible to local optima, they propose an enhanced Splitting-EM algorithm based on Bayesian Ying-Yang learning [10]. This approach facilitates the training of Gaussian Mixture Models (GMMs) by dynamically determining the optimal number of Gaussian components. By establishing a unique,
user-dependent signature model, they ensure improved approximation accuracy. Experimental results demonstrate that employing wavelet packet analysis for feature extraction and training GMMs using the Splitting-EM algorithm achieves a verification accuracy of 95.8%, indicating a satisfactory outcome for signature verification. Krish et al. [11] explore signature verification on smartphones. They conducted their analysis on a database comprising 25 users and a total of 500 signatures obtained from Samsung Galaxy Note devices. Their verification algorithm integrates two approaches: one based on features (utilizing Mahalanobis distance), and the other on functions, employing Dynamic Time Warping (DTW). The authors reported an EER of 0.525%.

2.2. Bimodal Systems

Buriro et al. [7] extend “Touchstroke” [8] and leverage two human behaviors: how users hold the phone, and how they enter a 4-digit text-independent PIN/password. The scheme exploits built-in smartphone sensors (orientation, gravity, magnetometer, gyroscope, and accelerometer) to compute phone-holding behavior. Users, involved in the study, were allowed to enter any combination of 4-digit numbers and/or alphabets, making it comfortable to use. Experiments confirmed that every user has a unique phone movement behavior and a different way of touch-typing. The authors achieved as high as 99% True Accept Rate (TAR) at a False Accept Rate (FAR) as high as 5%.

The study by Shen et al. [12] examines the feasibility and applicability of using motion-sensor-based data for user authentication on smartphones. The study extracts features from sensory data to accurately characterize users’ passcode-input actions and applies one-class learning methods for user authentication. The study includes experiments with data from 48 participants and 129,621 samples across various operational scenarios and different types of smartphones. The results showed an FRR of 6.85% and an FAR of 5.01%. The dataset used in the study is publicly available for future research.

Our bi-modal authentication approach represents a significant advancement in smartphone security by leveraging touch-typing and signing behaviors. Unlike older techniques that often rely on single modalities or static features, our approach combines dynamic touch data with comprehensive profiling to create robust user templates. By capturing touch interactions, such as finger movement velocity and touch timings, we enhance authentication accuracy while maintaining user convenience. Compared to traditional methods, which may suffer from limited adaptability or susceptibility to spoofing, our approach offers a more holistic and adaptable solution. Through rigorous experimentation and evaluation, we have demonstrated its efficacy in achieving high levels of accuracy and resilience to fraudulent attempts. As mobile devices continue to evolve, our approach stands poised to meet the increasing demand for secure and user-friendly authentication methods in the ever-expanding digital landscape.

3. Our Approach

Our approach is a bi-modal authentication system that exploits the smartphone’s built-in touchscreen to collect touch features generated during the entire course of touch-typing of an 8-digit “text-independent” number and signing. It records all touch-points and the velocity of
finger movements during this process, as well as touch-timings extracted during the keystroke entry. All these unimodal profiles are preprocessed to extract productive and useful features. These extracted features are then used to create the user profile template and are stored in the database. Our approach matches every testing sample against the stored user templates and authenticates or rejects based on their similarity. The whole procedure is illustrated in Figure 1.

3.1. Data Source
We utilized the touchscreen as our primary data source, considering it the main user interface for the device, categorized into single or multi-touch functionalities. The Android library MotionEvent [13] provides a class for tracking the motion of different pointers, such as a finger or stylus, and reports touch events using an object of this class, which contains information about the location, pressure, size, and orientation of the touch. Moreover, the Android VelocityTracker class is employed to monitor the motion of the pointer on the touchscreen and acquire velocities in the X and Y axes. We relied on this class to collect touch-points generated from finger-based signature acquisition.

3.2. Success Metric
We use the following metric to report our obtained results:

- True Acceptance Rate (TAR): The TAR represents the ratio of correctly classified “owner” samples to all the classification attempts for “owner” samples.
- **False Reject Rate (FRR):** FRR indicates the ratio of “owner” attempts misclassified as “non-owner” samples by the system. It can be computed as the complement of the TAR.
- **False Accept Rate (FAR):** FAR represents the ratio of incorrectly classified “non-owner” samples as “owner” samples. It can be computed as the complement of the True Reject Rate (TRR).
- **True Reject Rate (TRR):** TRR is the ratio of correctly classified “non-owner” samples as “non-owner” by the system.
- **Accuracy:** The Accuracy represents the overall effectiveness of the system in correctly classifying both “owner” and “non-owner” samples.

4. **Methodology**

In this section, we explain the procedure of our analysis:

4.1. **Data Collection**

We developed an Android application, STHAuth, which can be installed on any Android smartphone starting from version 8.0.4. We recruited 20 volunteers, comprising a mix of males and females. Participants were primarily Master’s or Ph.D. students but not security experts. The volunteers were diverse in terms of nationality. The purpose of the experiment and the proposed solution were explained to each volunteer, and they provided explicit consent to participate. We collected data in three postures: sitting, standing, and walking with a Samsung Galaxy S21 5G. We utilized the interface of our data collection shown in Figure 2. We separately collected training and testing samples from each of the recruited users. Specifically, we collected 30 training and 10 testing samples from each user. Thus, we managed to collect a total of 800 samples for this analysis.

4.2. **Feature Extraction**

We captured the way the user inserts a PIN of predefined length using a 10-digit numpad. In this context, a full keystroke gesture is recorded from the moment the user presses the first digit to the last one with regards to the currently selected PIN length [7]. To this end, we extract the following features from 8-digit “free-text” touch-strokes: (i) the duration of each stroke (D1 to D8), (ii) interval duration between strokes (F1type1 to F8type1), (iii) interval duration between the end of each stroke (F1type2 to F8type2), (iv) the interval duration between the start of successive keystrokes (F1Type3 to F8Type3), and finally (v) the full duration of the keystroke sequence (FType4). Hence, the final feature vector for each of the touchstroke samples is 30 features long.

Similarly, to register the signature fingerprint of each user, similar to [14], we extracted the following features from the signature drawn by the user: (i) Start and end values for the X and Y coordinates, (ii) Euclidean distance between the start and end points, (iii) standard deviation related to the set X and Y values associated with the points making up the signature, (iv) X and Y values related to the area spanned by the signature, and (v) maximum and average velocity values for the X and Y coordinates (These values are obtained from the set of velocities measured...
between the individual points making up the signature gesture. Additionally, we segmented the entire signature into 10 parts and extracted segmentation-based X and Y features, as well. The final feature vector for this sample is 33 features long.

4.3. Feature Fusion

The study [15] explains different levels of data fusion. The study suggest to fuse extracted data as early as possible. We fuse the data at feature level as this level aims to provide the most relevant information to the recognition system. As such, we concatenate features of individual modality (touchstroke and signature) and form a final feature vector of 63-feature long.

4.4. Classifiers Selection

The proposed framework aims to address the challenge of user authentication within client-server architectures, such as those used in banking and remote access systems. In this context, we trained our chosen classifiers on samples from multiple users. We chose 3 logically different classifiers, namely, KNN, RF, DNN in our analysis.

t-Distributed Stochastic Neighbor Embedding (t-SNE), introduced by van der Maaten and Hinton in 2008 [16], serves as a powerful tool for visualizing high-dimensional data in a lower-dimensional space, typically two-dimensional. Unlike linear techniques such as PCA, t-SNE is nonlinear in nature, allowing it to capture intricate relationships and structures within the data. Its strength lies in its ability to preserve both local and global structures: local structures pertain to the relationships between neighboring data points, while global structures encompass relationships across the entire dataset. This versatility makes t-SNE an invaluable asset for
exploring and understanding complex datasets, offering insights that linear methods might overlook.

The decision boundaries shown in the t-SNE plots (see Figure 3) provide insightful information on how our classifier is making sense of the users’ data. The way data points are spread out and clustered together in the t-SNE space depicts a real sense of confidence in our chosen classifier’s ability to understand the nuances of our dataset. We can see clear boundaries between different clusters, indicating that the classifier has picked up on meaningful patterns and can reliably distinguish between different classes. This tells us that our classifier isn’t just memorizing the training data but is actually learning to generalize well to new, unseen examples. These t-SNE plots offer a compelling visual validation of the classifier’s performance and reinforce our trust in its ability to accurately classify data based on the features it has learned.

4.5. Parameter Optimization

Parameter optimization of classifiers involves fine-tuning the parameters of chosen algorithms to achieve optimal performance. This typically involves techniques such as grid search or random search to systematically explore the hyper-parameter space. The goal is to find the best combination of hyperparameters that could potentially maximize the model’s performance metric. Parameter optimization is crucial for optimizing the predictive power and generalization ability of machine learning models.

To achieve this, we conducted a search to determine the optimal number of layers, the
Table 1
Parameter optimization of all chosen classifiers. The parameters “# of Neighbors” is for KNN and “C” and “γ” are regularization parameters for SVM, respectively.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Range of Parameters</th>
<th>Best Value</th>
<th>Best validation Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td># of Neighbors</td>
<td>1</td>
<td>94.27</td>
</tr>
<tr>
<td></td>
<td>“C”</td>
<td>0.1 to 5.0 (stepsize=2)</td>
<td>2.3</td>
</tr>
<tr>
<td>SVM</td>
<td>“γ”</td>
<td>“1″, “0.1″, “0.01″, “0.001″</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>“Kernel”</td>
<td>“linear”, “rbf”, “poly”, “sigmoid”</td>
<td>linear</td>
</tr>
<tr>
<td>DNN</td>
<td>“num_layers”</td>
<td>2 to 10 (stepsize=1)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>“num_units”</td>
<td>32 to 512 (stepsize=32)</td>
<td>288, 512, 384</td>
</tr>
<tr>
<td></td>
<td>“learning_rate”</td>
<td>“0.01″, “0.001″, “0.0001″</td>
<td>0.001</td>
</tr>
</tbody>
</table>

required units in each layer, and the appropriate learning rate. It’s important to highlight that we conducted a grid search to identify the best hyperparameters of the adopted classifiers from the parameters listed in Table 1. We utilized Scikit-learn\(^3\), a Python library, for hyperparameter optimization of baseline classifiers, and Keras-tuner\(^4\) for the DNN network. Our training set was utilized to optimize the chosen classifiers, ensuring that the test set remained unseen during this process. Subsequently, we trained the optimized classifiers and evaluated them on our previously separated test set to report our final results.

5. Results

We summarise our obtained results in terms of TAR, FRR, FAR, TRR, and Accuracy. Since the FRR and TRR can be computed as \(1 - TAR\) and \(1 - FAR\), respectively, we report only TAR and FAR to avoid any redundancy.

In Figure 4, we present an overview of the outcomes yielded by our carefully optimized classifiers when trained on both individual (unimodal) and combined (bimodal) datasets. The corresponding visualizations are showcased in Figures 4a and 4b. Our objective in this comparison was to assess the impact of each modality independently; notably, the signature patterns demonstrated notably higher inter-class variability (refer to Figure 4b). Subsequently, we merged the two modalities and trained our refined classifiers on the fused dataset, leading to the outcomes illustrated in Figure 4c. It’s noteworthy that across all configurations, the performance of the DNN remained consistently superior to that of other classifiers.

Achieving a maximum TAR of 97.1% and a FRR of 2.9% in bimodal settings indicate a robust authentication system with high accuracy in correctly identifying legitimate users (see Figure 4c). While the FRR may appear slightly higher, it’s important to note the significant impact of the low FAR of 0.2%, highlighting the system’s resilience against unauthorized access attempts. This balance between high TAR and low FAR is crucial for maintaining security while ensuring

\(^3\)https://scikit-learn.org/stable/
\(^4\)https://keras.io/keras_tuner/
smooth user experience, as it minimizes the risk of unauthorized access without overly burdening users with false rejections. Overall, these performance metrics reflect a well-optimized authentication system with strong usability and security measures in place.

From Figure 4, it is evident that the DNN as a classifier outperformed other classifiers like KNN and SVM possibly because of the following reasons: Firstly, DNNs possess an intrinsic capability to learn well from the representations of data, enabling them to effectively capture complex patterns and relationships inherent in both low and high-dimensional datasets. This inherent capacity lends DNNs a distinct advantage in scenarios where the underlying data distribution is nonlinear or exhibits high-dimensional complexity, circumstances where linear models like KNN and SVM may falter due to their reliance on predetermined distance metrics or linear decision boundaries. Secondly, the expressive power of DNNs, because of their deep architecture, affords them the capacity to capture smaller variations and subtle dependencies present within the data. Finally, the optimization process inherent to DNN training, typically
involving backpropagation and stochastic gradient descent, engenders iterative refinement of model parameters, thus enabling DNNs to continuously improve their predictive performance over successive iterations. This iterative learning paradigm powers DNNs with a dynamic adaptability to the data distribution, enhancing their generalization capabilities beyond that achievable by KNN and SVM, enhancing their generalization capabilities beyond that achievable by KNN and SVM.

6. Conclusions

In this paper, we have introduced a bimodal behavioral biometric-based authentication scheme for smartphones. Our proposed scheme harnesses two distinct user behaviors – keystroke timings generated from entering an 8-digit "text-independent" PIN and user finger-tip written signature – for user registration and subsequent authentication. We emphasize the user-friendliness of our approach because it leverages users’ familiarity with both modalities. Moreover, the fusion of these two inherently secure behaviors significantly enhances the scheme’s security, as it becomes exceedingly challenging to replicate both keystroke timings and touch features.

Through comprehensive evaluation, comparing against four state-of-the-art machine learning classifiers (including KNN and SVM as baseline classifiers), our results unequivocally demonstrate the superior performance of DNN, achieving the highest accuracy rates of 98.45% when trained in the bimodal settings.

Our proposed system yields higher accuracy, serving not only as a standalone authentication method for smartphones but also as a complementary tool for existing methods like face recognition on laptops and fingerprint recognition in automotive applications. As future work, we plan to prototype a proof-of-concept application for cross-vendor platforms, involving a large number of participants, and conducting thorough evaluations from performance (power consumption, memory usage, and timing metrics), robustness to attacks (random, mimic, engineered), and usability perspective.

Acknowledgments

Authors would like to thank all the volunteers for their participation in the experiments and groupmates for their valuable and insightful thoughts.

This work is partially supported by projects “SEcurity and RIghts In the CyberSpace - SERICS” (PE00000014 - CUP H73C2200089001), “Interconnected Nord-Est Innovation Ecosystem - iNEST” (ECS0000043 - CUP H43C22000540006), and PRIN/PNRR “Automatic Modelling and Verifi-
cation of Dedicated sEcUrity deviceS - AM∀DEUS” (P2022EPPHM - CUP H53D23008130001), all under the National Recovery and Resilience Plan (NRRP) funded by the European Union - NextGenerationEU.

References

[1] B. Spencer, Mobile users can’t leave their phone alone for six minutes and check it up to 150 times a day, 2013. URL: https://www.dailymail.co.uk/news/article-2276752/Mobile-users-leave-phone-minutes-check-150-times-day.html.


