Age, Gender and Operating-hand Estimation on Smart Mobile Devices

Attaullah Buriro*, Zahid Akhtar‡, Bruno Crispo†* and Filippo Del Frari*
*Department of Information Engineering and Computer Science (DISI),
University of Trento, Via Sommarive, 38123, Italy,
Email*:{attaullah.buriro, bruno.crispo, filippo.delfrari}@unitn.it
†DistrNet, KULeuven, Belgium
Email†: bruno.crispo@cs.kuleuven.be
‡Department of Mathematics and Computer Science, University of Udine, Italy,
Email‡: zahid.akhtar@uniud.it

Abstract— **In recent years, studies have explored the possibility of extracting ancillary information, such as gender, age, height, weight, etc., from primary biometric traits, namely, iris, fingerprint and face, however, the estimation of soft biometrics from mobile biometric data associated with behavioral modalities, e.g., touch and phone movement, is still a much less explored area. This paper investigates the possibility to estimate soft attributes on smart mobile devices. In particular, we design a scheme to estimate age, gender, and operating-hand using information originated from keystrokes, when the user enters her secret PIN/password. The experimental analysis of the devised scheme on the publicly available keystroke dataset 'TDAS' shows promising results. The proposed method attained the highest accuracy of 87.7%, 82.8%, 95.5%, respectively, for age, gender, and operating-hand estimation via timing-based keystroke features.**

Index Terms—**Biometrics, Authentication, Human-Computer Interaction**

I. Introduction

In recent years, studies have shown that ancillary information, such as gender, age, height, weight, etc., can be extracted from primary biometric traits such as iris, fingerprint, and face [1]. This ancillary information is also known as soft biometrics. The recognition accuracy of biometric systems based on primary traits can be improved by incorporating the user's ancillary information [1]. There exist several studies that have explored well-adopted physical biometrics for this purpose [2]. However, very few works have been carried out investigating the possibility of estimating soft biometrics attributes from behavioral biometrics collected on mobile devices. In the keystroke-based authentication system, the straightforward method to obtain soft biometrics information is asking users to explicitly provide such attributes at the time of enrollment. However, this may annoy the user having a negative impact on usability. Therefore, another way is estimating these attributes from the keystroke timings during authentication.

Smartphones have become pervasive. They are being used for many applications besides making and receiving phone calls. The estimation of user's soft biometrics attributes on mobile devices has a lot of potential applications. For instance, soft biometrics could be used: (i) to improve the user interactions (e.g., themes can be customized automatically), (ii) for age-specific access control (e.g., under aged children can be prevented from watching certain movies or accessing certain websites), and, (iii) for effective gender-age-specific product advertisements or suggestion for new applications, etc.

This paper focuses on estimating soft attributes (gender, operating-hand, and typist age) based on keystroke timings on mobile devices. The main contributions of this work are as follows:

- A novel method to estimate gender, operating-hand, and age of the user, when she types their 4 to 16 digit long PIN/password.
- The use of different classifiers, among which an unsupervised AutoEncoder (AE) neural network, for soft biometrics, attributes estimation from 4/6/8/10/12/14/16-digit text-typing.
- The validation of the efficacy of the proposed mechanism on a dataset of 150 users.

The paper is organized as follows. Section II presents the related work. The proposed approach for gender, operating hand and age estimation is described in Section III. Experimental protocol, dataset, results, and analysis are discussed in Section V. Section VII concludes the paper.

II. Related Work

Estimation of soft-attributes has been an interesting topic with a lot of potential applications, besides increasing the accuracy of user authentication. Some of the applications include target advertisement over social networks, digital forensic analysis, and improvement of user interaction.

Jain *et. al* [2] showed that the accuracy of a fingerprint recognition system can be improved by using soft-attributes such as ethnicity, gender, and height. However, their approach requires user cooperation and additional equipment to measure the height, thus not suitable for smartphones. Keystroke dynamics is the most suited behavioral modality to extract soft-attributes both on computers and smartphones, since it can be collected unobtrusively, and requires no extra equipment. In fact, estimation of soft-attributes on *personal computers* using keystroke dynamics has been largely investigated [3]–[5]. In particular, Idrus *et al.* [4] presented an

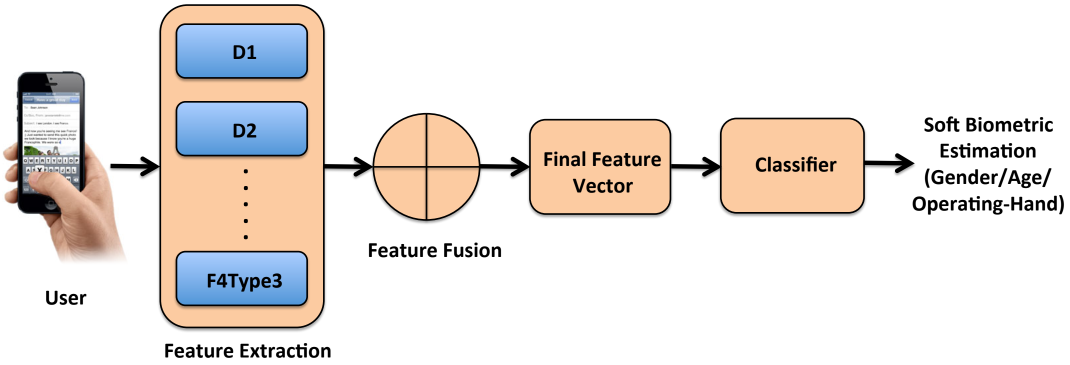


Fig. 1: The steps involved in soft biometrics estimation on mobile devices.

algorithm to estimate four soft-attributes from the entered keystrokes. They estimated gender, age, operating-hand, and whether the text was typed with one or two hands. The experiments were conducted on a private dataset of 110 users. The dataset includes 78 males and 32 females, 98 right and 12 left-handers with the age of the users ranging from 15 to 65 years. The users were classified into two classes, i.e., class 1 containing users of age ≤ 30 years, and class 2 containing users of age ≥ 30 years. The timing-based features from keystrokes of different key-lengths (17-24) were extracted to be fed to the LIBSVM classifier. The reported accuracy of gender, age and operating-hand estimation using different lengths of entered text are 65%-90%, 65%-82%, and 70%-90%, respectively.

Very few studies have been conducted for soft biometrics estimation on smartphones [6], [7]. The authors in [7] investigated the feasibility of estimating gender from the touchscreen swipe gestures on smartphones. They used the dataset created under the SuperIdentity project [8]. The multimodal dataset consists of 116 users (57 males and 59 females) with face, iris, voice, keystroke, swipe, signature, gait, hand and fingerprint modalities. The swipe data was collected on Samsung (GT-I9100) 'Galaxy S2' smartphone with 4.3" capacitive touchscreen. A voting based scheme using four well-known classifiers (i.e., Naive Bayes, Logistic Regression, Support Vector Machine and Decision Trees) was used achieving the maximum accuracy of 78%.

Existing works on soft biometrics estimation on mobile devices either estimate a single attribute (e.g., gender in [6], [7]) or utilize data from multiple sensors (e.g., use of accelerometer and gyroscope sensors in [6]). While, in this paper, we estimate three soft-attributes on smartphones using only a single modality, i.e., keystroke dynamics.

III. Soft Biometrics Estimation on Mobile Devices

Keystroke dynamics on smartphones has been well explored in recent years. Used as biometric authentication, it provides an additional security layer to the existing access control methods. When the user enters her PIN/password, the identity of the user is confirmed by matching different keystroke timings. However, a little research has been carried out to infer other associated user attributes, such as age, gender, and so on. In this work, we design an algorithm to estimate age, gender, and operating-hand via information originating from the keystrokes, when the user enters her secret PIN/password. We treat gender and operating-hand as a binary classification, and age-estimation as 3-class classification problem as per World Health Organization (WHO) factsheet[1]. We define class 1 as teenagers ($<$20), class 2 as adults ($\geq 20 < 60$), and class 3 as senior citizens (≥ 60).

The proposed framework first extract timing-based keystroke feature vectors from the entered text. The vectors in the feature space are then fed to a specific classification scheme that includes Naive Bayes, Support Vector Machine, Random Forest and AutoEncoder based Deep Neural Network, which determines whether the feature description corresponds to the true class or not. Figure 1 illustrates the steps involved in our process.

A. *Classification Methods*

In this study, we have applied a set of 5 logically different machine learning classifiers. The classifier selection depends on various parameters such as simplicity, the size of the dataset, linearity/non-linearity of the dataset, training time, and computational constraints, etc. It is usually impossible to understand in advance which classifier fits a dataset better. Our classifier toolbox consists of simple and advance state-of-the-art machine learning classifiers, namely, Naive Bayes, Support Vector Machine(SVM), NeuralNet, RandomForest (RF) and Deep Neural Network (DNN). All the classifiers except the DNN, are considered useful for both small and larger datasets, hence we applied 50% for training and remaining 50% for testing them. However, for DNN, we have tried 50% and 80% for training and remaining 50% and 20% for testing the classifier, simulataneously. We used PRTools, a Matlab-based toolbox, for all the adopted classifiers except for the DNN, which was implemented from standard Matlab 2016a workbench.

IV. Experiments

In this section, we provide an experimental evaluation of the proposed soft biometrics estimation framework. soft-

[1] http://www.who.int/mediacentre/factsheets/fs334/en/

attributes framework.

A. Dataset

Though, there exist several datasets reporting keystroke dynamics data, the only we found that is publicly available and contains the participant's soft biometrics information is the 'TDAS' [9] dataset. Therefore, we used this dataset as ground truth to evaluate our framework. The dataset is composed of 150 participants with 10 samples from each participant. A diverse range of profession and age (covering 6 age-groups, i.e., 10+, 20+, 30+, 40+, 50+, and 60+) among participants are present. The dataset was collected on a $10.1''$ wide Samsung Galaxy Tab (GT-P7510) digital tablet powered by 1GHz dual-core processor and equipped with a 1-GB RAM with collection tool using Android API level 15. It is worth noting that the dataset was collected in different user-tagged locations, i.e., in their offices, cars, etc. The datset includes keystroke timings plus the touch pressure and size for two numerical lengths, i.e., 4-digit (e.g., '5560') and 16-digit (e.g., '$1379666624680852'$).

B. Experimental Protocol

We randomly selected 50% of the data samples as a training set, whereas the remaining 50% data samples were used to build the testing set. The 'TDAS' database is a highly unbalanced dataset: Male (45) and Female (105), Right-Hander (136) and Left-Handers (14), hence the computed results might get influenced by the frequently occurring class. To overcome this limitation, we utilized the Synthetic Minority Oversampling Technique (SMOTE) [10], which resamples the original dataset to increase the samples of minority class to make it equal to the majority class. For instance, the dataset contains 450 and 1050 samples for the male and female classes, respectively. By using Weka's filter with 133% SMOTE, we increased the number of male class samples up to 1048 samples, which is closer to 1050 samples of female class. Similarly, for hand and age datasets, we increased the number of samples of minority classes. In age dataset, the majority of participants are from 20+ age class, therefore we took this class as the reference and applied SMOTE on all other age-groups to increase their number of samples.

Like in [11], we extracted timing-based keystroke features from the 16-digit entered text for different PIN/password lengths. The extracted feature vectors are of sizes 14, 22, 30, 38, 46, 54, and 62 for 4, 6, 8, 10, 12, 14, and 16 digit passwords, respectively.

We have used different feature lengths because the different application requires the different size of PIN/passwords. For instance, 4-digit is used often for one-shot login, 8-digit length is common for emails, facebook, twitter, linkedin, and so on.

V. Results

We report our obtained results in terms of True Acceptance Rate (TAR), False Acceptance Rate (FAR), False Rejection Rate (FRR), accuracy and DET curves. Accuracy is the ratio of number of correct decisions to the number of all decisions. DET is used to show the correlation between the two common error types, i.e., FAR on the x-axis and FRR on the y-axis. The curve closer to (0,0) coordinates indicate better performance.

In figures 2 and 3, we report the accuracy results attained by the proposed algorithm using all classifiers for the mentioned lengths. Only, the trade-off between the FRR and the FAR errors are depicted, and the trade-off between the TAR and the and FAR are not reported in order to avoid redundancy; since the TAR and TRR can be estimated by computing $1 - FRR$ and $1 - FAR$, respectively. Similarly, the classification error can be computed as $100\% - accuracy$.

It can be seen from the figures 2a and 3a that Naive Bayes and DNN did not perform well on this dataset for gender estimation. Namely, the maximum accuracies reached by NB and DNN are 63.9% and 56.4% (57.9% for 80/20[2] split), respectively. Also, it is easy to see that the Random Forest (RF) classifier outperformed other classification schemes for gender classification. The maximum accuracy obtained by RF is 82.8% over 14-digits.

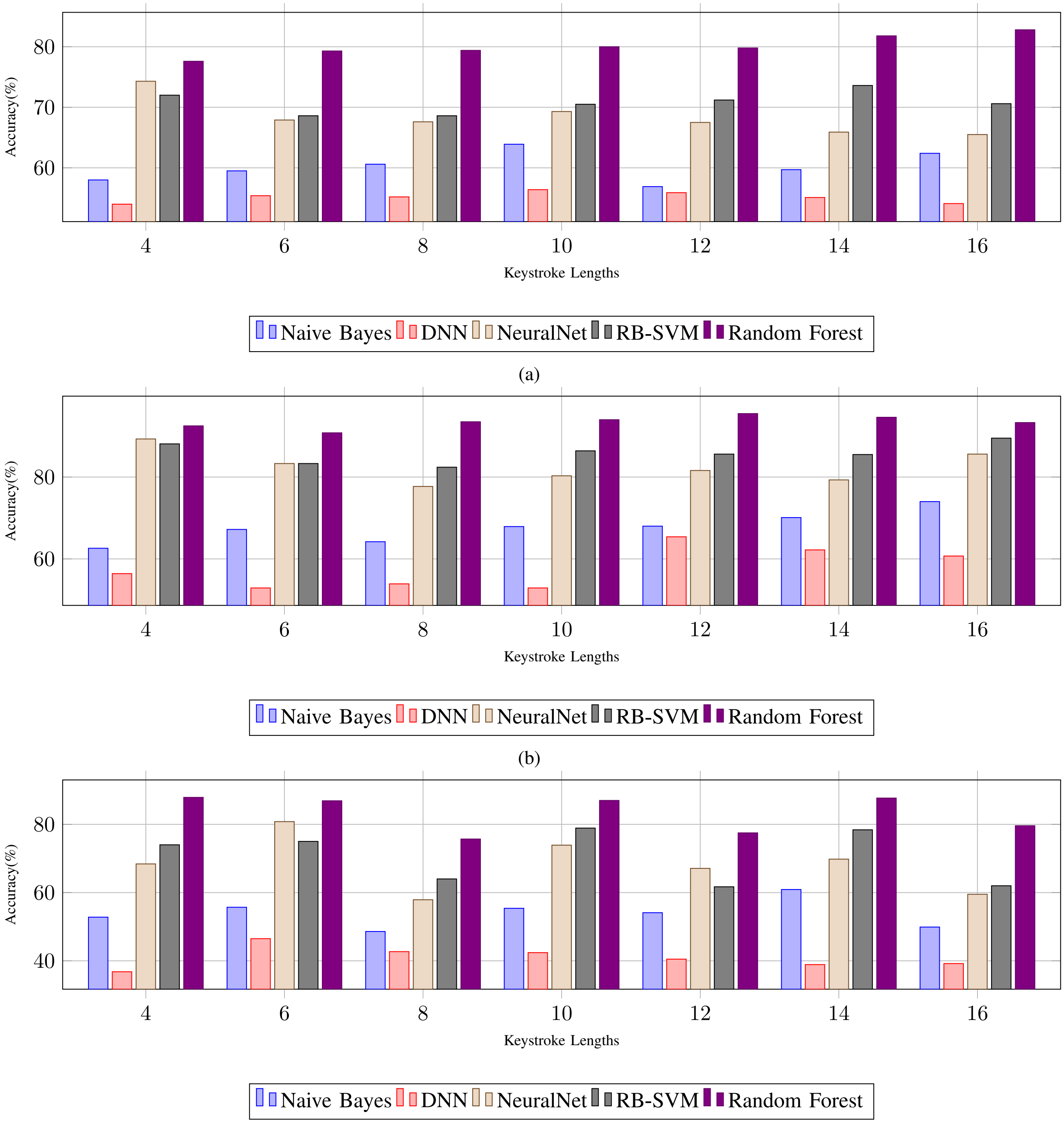
The results of operating-hand estimation are illustrated in the figures 2b and 3b. Especially, the bar graph (Figure 2b) shows the accuracy of different classifiers for different keystroke lengths. The highest accuracies achieved by using NB, DNN, NeuralNet, SVM and Random Forest classifiers are 74% (14-digit), 65.4% (12-digit, compare to 56.2% for the same length with 80/20 split), 89.3% (4-digit), 89.5% (16-digit) and 95.5% (12-digit), respectively. We can observe that the Random Forest classifier, like for gender classification, performed well for operating-hand estimation, as well.

The age estimation results are depicted in the figures 2c and 3c. We can observe that for age estimation, DNN performs worst among the adopted classifiers in this work. The highest accuracy obtained by DNN is 42.5% (compare to 51.1% for 80/20 split for the same length). While, the highest accuracies achieved with NB, NeuralNet, SVM and Random Forest classifiers are 60.9% (14-digit), 80.8% (6-digit), 78.9% (10-digit) and 87.9% (4-digit), respectively. It is worth mentioning that authors in [4] formulated the age estimation as a binary classification problem (class 1 <30 years, and class 2 $\geq$ 30 years), and obtained maximum accuracy of 82%. Whereas, in this work, we formulated the age estimation problem as 3-class classification problem, and attained the maximum accuracy of 87.9%. The highest accuracies were reached by Random Forest classifier with 4-digits, 10-digits and 14-digits, however due to space limiations, we show DET curve for only 4-digits length (see Figure 3).

VI. Discussion on Results

Experimental results on the publicly available 'TDAS' database are promising. They suggest that the proposed method attains better performance than existing techniques using the Random Forest (RF) classifiers. In particular, the

[2] 80% training and 20% testing

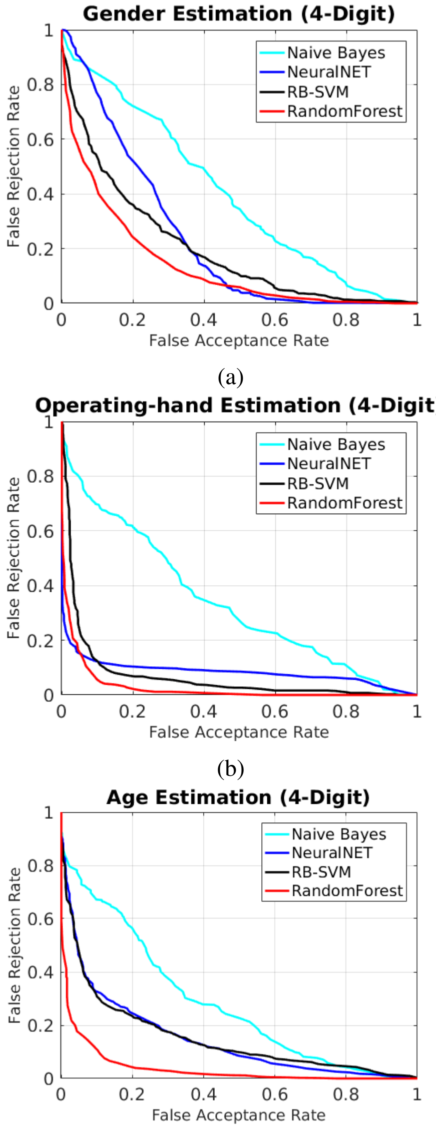


(c)

Fig. 2: Bar graphs for (a) gender, (b) operating-hand, and (c) age estimation for different keystroke lengths obtained from all the classifiers

RF method obtained 82.8% accuracy for gender recognition (as compared to 60.3% in [12]), 95.5% for operating-handedness (as compared to 70% to 90% in [4]), and 87.9% for age estimation (as compared to 82% in [4]). The RF classifier provide the highest accuracy for gender, operating-hand and age estimation at 16-digit (see Figure 2a), 12-digit (see Figure 2b), and 4-digit (see Figure 2c) long keystrokes. It is also worth noticing that the accuracy of most of the classifiers increases as the length of strokes increases.

Despite its simplicity, RF method outperforms other schemes under varying features and tasks owing to its ability to reduce the variances, averaging out the biases and the most unlikeliness of over-fitting. These statistical properties are mainly attained by the bagging. The performance of DNN is not so encouraging because of the scale of the dataset, namely it requires relatively large datasets to work well. We



(a)

(b)

(c)

Fig. 3: DET curves for (a) gender, (b) operating-hand, and (c) age-group estimation for all the classifiers for 4-digit length.

used this classifier in two settings: training with 50% and 80% and testing with remaining 50% and 20%, respectively, but the results are similar. Moreover, since the dataset used in this study has several outliers , which may be causing performance degradation in SVM and NeuralNet. Though NB enjoys the simplicity and computational efficiency, it substantially performed poorly. This may be occurring due to its assumption that all attributes are independent (i.e., no correlation between variables). Since it has been shown in several studies that correlation mapping is beneficial to attain better accuracy. In addition, NB also presumes that the data follow Gaussian distribution, which is generally not true in our case. Thus, NB either gets over fitted or fails to address the problem of concept-drift.

VII. Conclusions and Future work

This paper presents a solution for the estimation of soft attributes (i.e., gender, operating-handedness, and age) from the entered touchstrokes on smart mobile devices. The proposed method uses different classifier techniques.

Our proposed technique is simple, effective and transparent (estimates the attributes without letting the user know), which makes it highly suitable for real-time mobile devices.

As a part of future work, we investigate the use of more robust features. Also, we will consider ways to fuse the outputs of the soft biometrics estimator at the feature, score and decision levels in a systematic way with authentication schemes. Our aim is to assess if soft biometric can effectively improve authentication accuracy in the context of smartphones.

Acknowledgment

The authors would like to thank Prof. Ke Chen of the University of Manchester for providing TDAS dataset.This work was partially supported by the EIT Digital SecurePhone project and European Training Network for CyberSecurity (NeCS) grant number 675320.

References

[1] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[2] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Defense and Security*. International Society for Optics and Photonics, 2004, pp. 561–572.

[3] I. Tsimperidis, V. Katos, and N. Clarke, "Language-independent gender identification through keystroke analysis," *Information & Computer Security*, vol. 23, no. 3, pp. 286–301, 2015.

[4] S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and P. Bours, "Soft biometrics for keystroke dynamics," in *International Conference Image Analysis and Recognition*. Springer, 2013, pp. 11–18.

[5] R. Giot and C. Rosenberger, "A new soft biometric approach for keystroke dynamics based on gender recognition," *International Journal of Information Technology and Management*, vol. 11, no. 1-2, pp. 35–49, 2012.

[6] A. Jain and V. Kanhangad, "Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings."

[7] O. Miguel-Hurtado, S. V. Stevenage, C. Bevan, and R. Guest, "Predicting sex as a soft-biometrics from device interaction swipe gestures," *Pattern Recognition Letters*, vol. 79, pp. 44–51, 2016.

[8] S. Black, S. Creese, R. Guest, B. Pike, S. J. Saxby, D. Stanton Fraser, S. V. Stevenage, and M. Whittty, "Superidentity: Fusion of identity across real and cyber domains," *ID360: Global Identity*, 2012.

[9] P. S. Teh, P. S. Teh, N. Zhang, N. Zhang, A. B. J. Teoh, A. B. J. Teoh, K. Chen, and K. Chen, "Tdas: a touch dynamics based multi-factor authentication solution for mobile devices," *International Journal of Pervasive Computing and Communications*, vol. 12, no. 1, pp. 127–153, 2016.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] A. Buriro, B. Crispo, F. Del Frari, and K. Wrona, "Touchstroke: smartphone user authentication based on touch-typing biometrics," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 27–34.

[12] Typeguess: Using mobile typing dynamics to predict age, gender and number of fingers used for typing. http://cs229.stanford.edu/proj2014/Baris%20Akis,%20Mariano%20Sorgente,%20Abraham%20Starosta-Typeguess.pdf.