Annals of Emerging Technologies in Computing (AETiC)
Vol. 5, No. 5, 2021

Review Article

Identification of the Exclusivity of Individual's Typing Style Using Soft Biometric Elements

Mohd Noorulfakhri Yaacob[1*], Syed Zulkarnain Syed Idrus[1], Wan Azani Wan Mustafa[1], Mohd Aminudin Jamlos[1] and Mohd Helmy Abd Wahab[2]

[1]Universiti Malaysia Perlis, Malaysia
fakhri@unimap.edu.my; syzul@unimap.edu.my; wanazani@unimap.edu.my; mohdaminudin@unimap.edu.my
[2]Universiti Tun Hussein Onn Malaysia, Malaysia.
helmy@uthm.edu.my
*Correspondence: fakhri@unimap.edu.my

Received: 17th October 2020; Accepted: 15th November 2020; Published: 1st April 2021

Abstract: Biometric is used as a main security fence in a computer system. The unique characteristics of a person can be distinguished from each other. Human's biometrics can be categorized into three types: morphological, biological and behavioural. Morphological biometrics uses physical features for recognition. Biological biometrics used to identify user based on biological features. Behavioural biometrics such as gender, culture, height and weight can be used as an additional security measure within a system. These biometric behavioural features are also known as soft biometric. This study uses soft biometric elements (gender, culture, region of birth and educational level) in the keystroke dynamic study to distinguish typing patterns in each of these categories. The Support Vector Machine (SVM) classification method is used to perform this classification for soft biometric identification. The results of this study have shown that soft biometrics in keystroke dynamic can be used to distinguish group of individuals typing.

Keywords: ***Keystroke dynamics; Soft biometrics; Identification; Behavioural biometrics***

1. Introduction

Keystroke dynamics (KD) is a method of identifying users based on how the operator uses a keyboard to type [1, 2]. This method does not require high hardware investment but only involves changes in the system or application. The system which uses this method must record the time interval between the two letters that the user type. This KD can be categorized as behavioral biometrics. Generally, the recognition using KD or biometric behavior is less popular compared to other biometric methods such as the use of fingerprint, iris and DNA. Various recognition techniques have been used in KD studies to enable higher accuracy when using KD. The combination of KD recognition with the soft biometric features available to a person has been done in previous studies. The earlier study of this combination was done by Idrus, Cherrier [3] in 2012 by studying the use of hands when typing using one or both hands.

Soft biometrics is a technique that can be used for user recognition. Each individual has his or her own unique identity that can be distinguished from each other [4]. The soft biometric information of each individual is not sufficient to distinguish a person accurately, but it does enhance the ability to identify when combined with other biometrics [5].

In the 18th century, the soft biometric recognition method was first used by GALTON [6]. Their research used three main features of soft biometrics: anthropometric measurements (arm length),

Mohd Noorulfakhri Yaacob, Syed Zulkarnain Syed Idrus, Wan Azani Wan Mustafa, Mohd Aminudin Jamlos and Mohd Helmy Abd Wahab, "Identification of the Exclusivity of Individual's Typing Style Using Soft Biometric Elements", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 10-26, Vol. 4, No. 5, 1st April 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.05.002, Available: http://aetic.theiaer.org/archive/v5/v5n5/p2.html.

Scar effect and Mole, and Body shape. In 2001, soft biometric recognition techniques using the criteria of gender, race, eye colour and height were developed by Heckathorn, Broadhead [7].

Other researchers continue to conduct study using the soft biometric elements available to each individual using the elements such as mole, iris or eye retina [8], scar effect / tattoo [9], body figure [9], gender [10, 11], [12], ethnicity [13], eye color, height [12], weight, hair color [14], age, BMI, walking style [15], sitting style [16], eyebrow, blood type [17], heartbeat, talking style [18], vein image [19], facial shape [20], facial skin / figure, and ear shape [17]. The results of previous studies have found that these elements can be used to identify individual users. This study incorporates soft biometric elements (culture, gender, region of birth and CGPA) in the KD study.

2. Soft Biometrics Application for Keystroke Dynamics

The incorporation of soft biometric criteria in KD has been performed since 2011 to date. Various soft biometric criteria have been used in their study. The study on the integration of soft biometrics and KD was started in 2011 by Epp, Lippold [21]. Their research was on the combination of emotional elements as soft biometric features with KD. The results of the study using emotion elements (Confidence, Hesitance, Nervous, Relax, Sad, Fatigue, Anger, Joy and Happiness) had achieved an accuracy of between 77.4% and 87.8%. In the same year, Fairhurst and Da Costa-Abreu [22] conducted a study on the classification of gender by typing, male and female. The 10-fold cross-validation method was used to analyze the obtained data. The result acquired was 95% accuracy.

Later, a similar study to Fairhurst was carried out by Giot and Rosenberger [23] in 2012 which identified the gender based on typing. The results of this study ranged from 87.32% to 91.63%. Subsequently in 2013, the emotional stress aspect as a soft biometric feature was used in the joint study of KD [24]. The results of this study showed whether the user is in depression or not.

Similarly, a study conducted by Nahin, Alam [25] have shown that a user's emotions can be identified based on one's typing style. There are seven categories of emotions studied, namely anger, disgust, guilt, fear, joy, sad and shame. The results obtained was above 80% accuracy by classifying the users according to the emotions studied.

Bakhtiyari, Taghavi [26] has explored the use of emotional elements as a factor on how a user uses a keyboard, touch screen, and mouse. This research compared other normal methods regularly used to identify emotion such as Electroencephalography (EEG) machines, facial expression, voice and body language. The highest accuracy percentage obtained was 93.20%.

The combination of soft biometric and keystroke dynamics has attracted Idrus, Cherrier [27], [28] and Idrus [29] to conduct a study to identify the users based on gender, age, left or right hand and handedness. The best EER obtained from his study was 5.41% using the majority voting technique. Subsequently, in 2015, the study was continued by Idrus using the penalty combination and reward combination techniques to reduce the EER rate obtained in 2014 [30]. The EER results obtained for the reward combination are better than the penalty combination of 23.11%.

In 2016, a study on gender identification by typing was done by Antal and Nemes [31], but the aspect of their research was to identify gender typing using the touch screen. The results showed that the detection accuracy was 64.76% on the keystroke dataset and 57.16% on the touch screen. Also, in 2016, Idrus, Cherrier [32] conducted a study to classify typing in several soft biometric categories, namely gender, age range and handedness. The results showed an accuracy of 63% to 96%.

Latest KD studies related to soft biometric were conducted by Kołakowska [33]. They continue a study conducted by Nahin, Alam [25] which identify users' emotions during typing. Five emotions were studied in their research: Happiness, Boredom, Fear, Anger and Sadness. The results of their study concluded that the user's emotions influence a person typing at that time. However, a person's emotional control or personal strength will influence this type of typing. Katerina and Nicolaos [34] did the next KD-related study by studying KD with mouse and hand movements while using a computer. Table 1 illustrates the results and soft biometric elements used in previous studies.

Based on the previous studies, it is apparent that the method of using a keyboard can be distinguished using soft biometric elements. Various methods in utilizing soft biometric can be adapted in the industry by constantly checking the typing changes of the recorded actual system user, whether the user is authentic or not [35].

The research done in this paper incorporates four soft biometric elements in the use of KD. The soft biometric elements used in this study were cultures in Malaysia, gender, the region of birth in Malaysia and educational level.

Table 1. Results and soft biometric elements used in previous studies

Year	Author	Result	Soft Biometric	Element
2011	Epp, Lippold [21]	Accuracy 77.4% - 87.8%	Emotion	Confidence, Hesitance, Nervous, Relax, Sad, Fatigue, Anger, Joy, Happiness
2011	Fairhurst and Da Costa-Abreu [22]	Accuracy 95%	Gender	Male, Female
2012	Giot and Rosenberger [23]	Accuracy 87.32% - 91.63%	Gender	Male, Female
2013	Gunawardhane, De Silva [24]	Can Identify whether User is Stressful or Not	Emotion	Stress
2014	Nahin, Alam [25]	Accuracy 80%	Emotion	Anger, Joy, Disgust, Guilt, Fear, Shame
2014	Bakhtiyari, Taghavi [26]	Accuracy 93.2%	Gender	Male, Female
2014	Idrus, Cherrier [27]	EER 5.41%	Gender, Age, Handedness	Male, Female
2016	Antal and Nemes [31]	Accuracy 64.76%%	Gender	Male, Female
2018	Kołakowska [33]	Person's strength also affects the typing activity.	Emotion	-
2018	Katerina and Nicolaos [34]	Usage of mouse and Keystroke Dynamics	Hand Movement	Mouse Usage

3. Identification Approach

Research on authentication system using keystroke dynamics requires specific hardware and software to record information about each user's typing pattern. Computers that were used to record information on the typing style were equipped with special software. Other applications besides windows were terminated to prevent computer from being in normal condition. Each user was instructed to type several words using the software available in the provided computer. Each time or interval for each character typed by the user was recorded in a database in the application. Each interval between letters obtained during typing was stored in the database and later analysed. Each category of soft biometrics surveyed would involve two phases, namely training phase and testing phase.

Support Vector Machine (SVM) had been selected as a technique to analyse and classify the raw data obtained. User authentication accuracy rate was measured and calculated for each category of soft biometric involved through SVM methods. The software used to execute SVM is MATLAB. Figure 1 shows an overview of the methodology used to perform this classification.

3.1. Individual Profiles Based on The Way of Typing

Classification of how to type was executed based on four categories of soft biometrics which are culture (Malays, Chinese and Indians), gender, educational level (CGPA - Cumulative Grade Point Average) and region of birth. This classification aims to isolate how users use the keyboard of each category. There are two approaches used in KD study which is the study based on a free text or a fixed text. The free text analysis is based on time-lapse between two consecutive letters or better known as digraphs whereas fixed text, the entire time category recorded in a word by the user is compared against the previous recorded time of the respective user. For example, the user is directed to type multiple times text / password during the enrolment process. The time interval of typing for each letter in the corresponding sentence was recorded. After that, the comparison is made by the system if the user types the same sentence for the second time and afterward. However, this study focuses on fixed text approach.

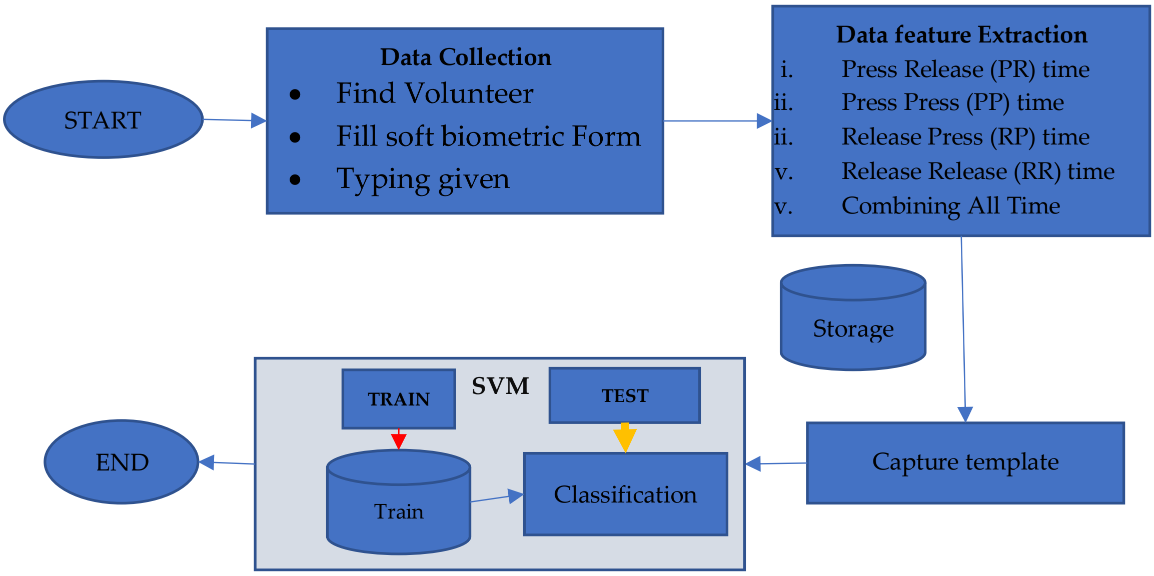

Figure 1. Methodology used to perform this classification.

3.2. Data Analysis

The analysis of keystroke data done in this study is based on four soft biometric criteria namely culture, region of birth, gender and educational level. The cultures to be studied are the three main residents in Malaysia namely Malay, Chinese and Indian [36], whereas the ROB category is divided into 4 parts: north of the peninsular, east of the peninsular, center of the peninsular and south of peninsular Malaysia. For measurements based on educational level, the CGPA result was used as two parts, 3.0 and above and 3.0 below. In addition, this study also incorporated other minority in Malaysia (Bajau, Murut, Siam, Suluk, Iban, Kadazan, Bisaya, Kedayan, Iranum, Tidong etc) and classified them into one group labelled as "Others". Hence, for the soft biometric gender category, the total number of categories are four, namely the three main races in Malaysia and one other category. Support Vector Machine (SVM) was used during the classification process.

The kernel used in this study is Radial Base Function (RBF)[37]. RBF kernel was selected due to its suitability for analyzing non-linear data recorded in keystroke dynamics [38]. This kernel is also able to isolate data to high dimension data. The data analyzed was separated into two parts, namely training data and test data. For example, if 1% of the total data analyzed is used as training data, then 99% is used as test data. This process is named as a training process within the SVM. This process was repeated 100 times starting from 1% training ratio up to 90% of the training ratio and the average was recorded. The classification of data for each category analyzed was labeled 1 and -1. For example, in the culture category, comparison of Malays and Chinese on their way of typing, Malays data was labeled as -1 and Chinese data was labeled as 1. This process was imposed on each of the two classes analyzed. The items analyzed are shown in Table 2 below.

Table 2. Soft biometric classification

Soft Biometric Category	Class 1	Class 2
Culture	Malays	Chinese
	Malays	Indians
	Indians	Chinese
	Others	Malays
	Others	Indians
	Others	Chinese
Region of birth	North of peninsular Malaysia	East of peninsular Malaysia
	North of peninsular Malaysia	Central of peninsular Malaysia
	North of peninsular Malaysia	South of peninsular Malaysia
	East of peninsular Malaysia	Central of peninsular Malaysia
	East of peninsular Malaysia	South of peninsular Malaysia
	Central of peninsular Malaysia	South of peninsular Malaysia
Educational level - CGPA	>=3.0	<3.0
Gender	Female	Male

This section clarifies the breakdown of statistical data collection obtained based on the four soft biometric criteria studied. Total number of volunteers involved was 250 people. Everyone was required to type 5 sentences given as many as 10 times correctly. Therefore, the total amount of keystroke data obtained in total is 12500. The statistical breakdown of the data is described in Figure 2 to Figure 5 below.

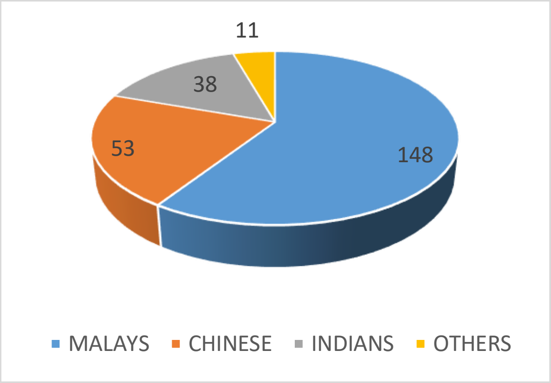

Figure 2. Shows the breakdown number of volunteers participated according to culture fragments

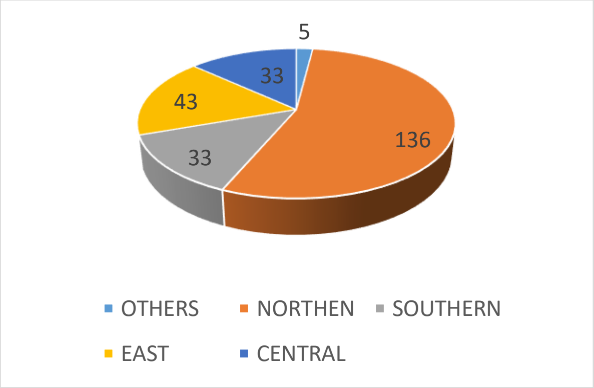

Figure 3. Shows the breakdown number of volunteers participated by region of birth fragments

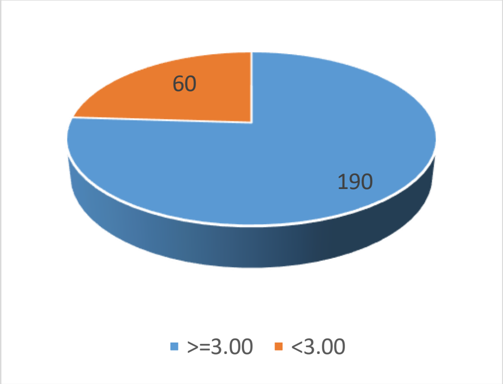

Figure 4. Shows the breakdown number of volunteers participated according to the CGPA fragments

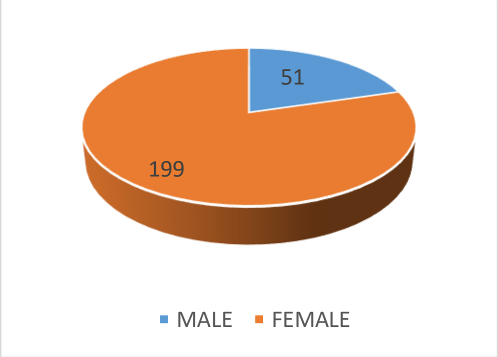

Figure 5. Shows the breakdown number of volunteers participated according to the Gender fragments

4. Experimental Results

This section describes the results of keystroke data analysis obtained.

4.1. Result Based on Culture

This study divides the culture into three main residents in Malaysia namely Malays, Chinese and Indians. Another minority group in Malaysia is in one category. Based on the entire data obtained (figure 2 - 5), the statistical breakdown of the data is given: Malays - 148; Chinese - 53; Indians - 38, Others - 11. According to the statistics of collected data, keystroke data from Indians was the lowest among the three main races. Malays made up the biggest number of volunteers, which is 148 personnel participated to contribute keystroke data. For analysis purposes, the number of participants of each culture was made equal for each category. For example, only 53 randomly selected Malays and all 53 Chinese records were used to compare these two classes. With regard to the Indians category, only 38 randomly selected records for each Malays or Chinese category were chosen for comparison with Indians. This is to balance the amount of data to be analyzed using SVM because according to a study by Idrus, Cherrier [27] , the number of data between two different classes should be equivalent to enable the best analysis results by SVM.

As explained in the previous chapter, all volunteers were required to type 5 sentences given as many as 10 times correctly. The first 3 of the 10 typing attempts were not analyzed to provide the volunteers opportunities to familiarize themselves with the word and sequence of letters in the given word. All five letters provided to users are simplified in Table 3. The total number of records analyzed for the class of each word is as below (Table 4).

Table 3: Texts used in this study

Abbreviation	Text / Password
P1	tunku abdul Rahman
P2	langkawi island
P3	english premier league
P4	instagram facebook twitter
P5	the sound of music

Table 4: Total number of Keystroke analysed for each class in culture

Class Category	Total typing	Total no of users for Each Class	Total Keystroke for Each Class
Malays VS Chinese	7	53	371
Malays VS Indians	7	38	266
Malays VS Others	7	11	77
Indians VS Chinese	7	38	266
Others VS Indians	7	11	77
Others VS Chinese	7	11	77

4.1.1. Chinese vs. Indians

Figure 6 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between Chinese and Indian using 5 words. A total of 38 samples of data typing volunteers from Chinese and 38 samples from Indian volunteers had been analyzed. The results obtained were quite good because of the 50% learning ratio, the accuracy earned had reached 75% for "instagram facebook twitter" and "the sound of music", while for the other 3 words the accuracy ranged from 78.5% to 83%. Table 5 shows a summary of the accuracy obtained from 50% to 90% of the learning ratio.

Table 5. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Chinese VS Indians

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	81.5%	82.6%	83.0%	83.3%	84.0%
P2	79.1%	81.1%	82.0%	83.4%	84.3%
P3	75.3%	77.5%	78.2%	79.4%	79.9%
P4	80.8%	81.6%	82.5%	83.5%	83.7%
P5	75.6%	76.6%	77.4%	78.8%	78.9%

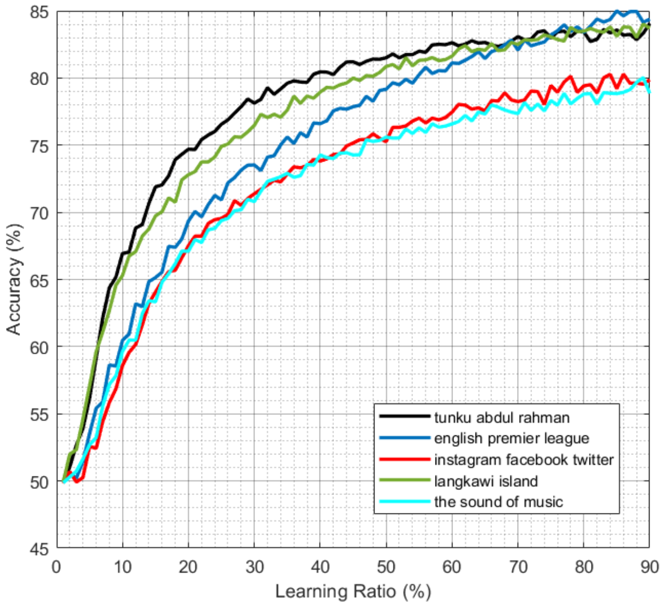

Figure 6. Average accuracy vs Learning Ratio (Chinese VS Indians)

4.1.2. Malays vs. Chinese

Figure 7 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the Malays and Chinese. The results obtained among the Malays and Chinese are better than the Chinese and Indians. Based on the 50% learning ratio, the accuracy obtained was between 83.5% and 88.5%. Table 6 shows a summary of the accuracy obtained for 50% of the learning ratio.

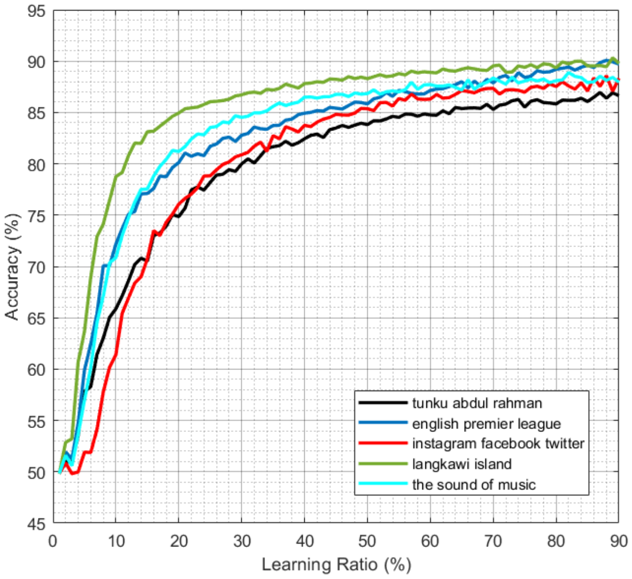

Figure 7. Average accuracy vs Learning Ratio (Malays VS Chinese)

Table 6. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Malays VS Chinese

Text	Leaning Ratio				
	50%	**60%**	**70%**	**80%**	**90%**
P1	83.8%	84.7%	85.3%	85.8%	86.2%
P2	88.3%	88.9%	89.5%	89.3%	90%
P3	85.8%	87.1%	87.8%	89.2%	90%
P4	85.4%	86.3%	87.3%	87.5%	88.2%
P5	86.8%	87.7%	88.4%	88.1%	88.2%

4.1.3. Malays vs. Indians

Figure 8 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the Malays and Indians. The results obtained between the two culture are 82% up to 88% accuracy for the learning ratio of 50% and above. Table 7 shows a summary of the accuracy obtained for 50% of the learning ratio.

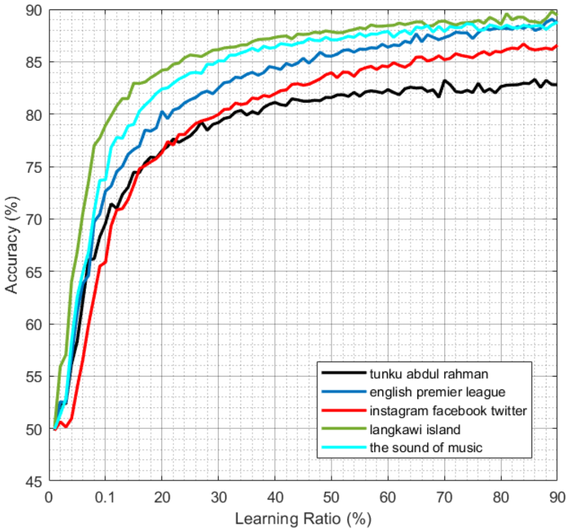

Figure 8. Average accuracy vs Learning Ratio (Malays VS Indians)

Table 7. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Malays VS Indians

Text	**Leaning Ratio**				
	50%	**60%**	**70%**	**80%**	**90%**
P1	81.6%	82.3%	83.2%	82.6%	82.8%
P2	87.8%	88.4%	88.8%	88.6%	89.4%
P3	85.5%	86.4%	87.3%	88.2%	88.8%
P4	83.9%	84.5%	85.2%	85.9%	86.6%
P5	87.0%	87.8%	87.8%	88.2%	88.8%

4.1.4. Others vs. Indians

Figure 9 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between Others and Indians. The results obtained between the two culture were 72.9% up to 90.4% accuracy for the learning ratio of 50% and above. Table 8 shows a summary of the accuracy obtained for 50% of the learning ratio.

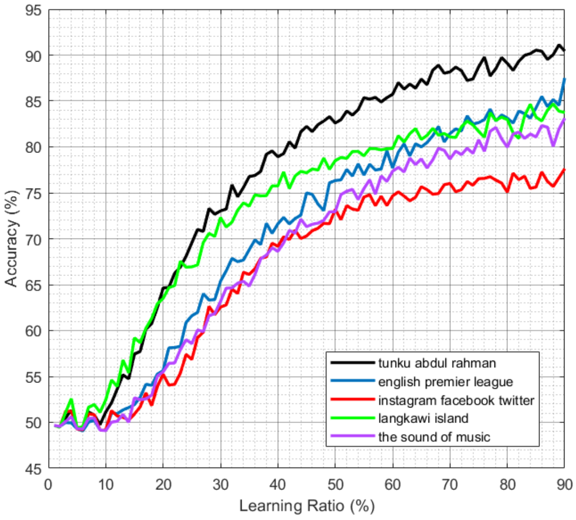

Figure 9. Average accuracy vs Learning Ratio (Others VS Indians)

Table 8. Summary of the accuracy obtained from 50% to 90% of the learning ration for Others VS Indians

Text	**Leaning Ratio**				
	50%	**60%**	**70%**	**80%**	**90%**
P1	82.6%	85.7%	88.2%	89.1%	90.4%
P2	78.5%	79.8%	81.1%	83%	83.8%
P3	76.3%	77.7%	81.1%	83.1%	87.5%
P4	73%	74.6%	76.0%	75.1%	77.6%
P5	73%	77.4%	78.7%	80%	83.1%

4.1.5. Others vs. Chinese

Figure 10 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between Others and Chinese. The results obtained between the two culture were 65.3% up to 81% accuracy for the learning ratio of 50% and above. Table 9 shows a summary of the accuracy obtained for 50% of the learning ratio.

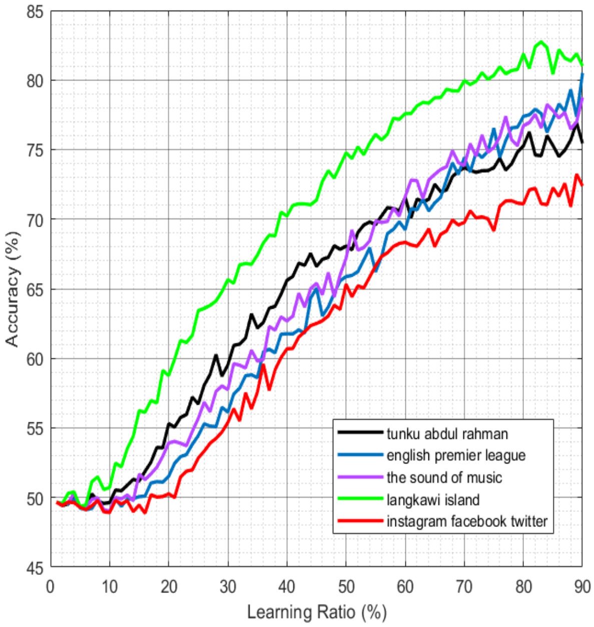

Figure 10. Average accuracy vs Learning Ratio (Others VS Chinese)

Table 9. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Others VS Chinese

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	68%	71.2%	73.7%	75.3%	75.4%
P2	74.8%	77.6%	80%	81.9%	81%
P3	65.9%	69.2%	74.4%	77.4%	80.1%
P4	65.3%	68.3%	69.8%	71.1%	72.4%
P5	67.2%	71.2%	73.8%	76.7%	78.7%

4.1.6. Others vs. Malays

Figure 11 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between Others and Malays. The results obtained between the two culture were 67.8%% up to 81% accuracy for the learning ratio of 50% and above. Table 10 shows a summary of the accuracy obtained for 50% of the learning ratio.

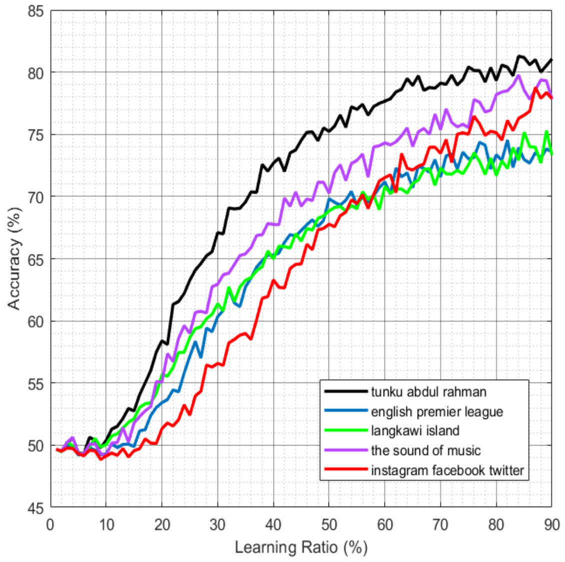

Figure 11. Average accuracy vs Learning Ratio (Others VS Malays)

Table 10. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Others VS Chinese

Text	**Leaning Ratio**				
	50%	**60%**	**70%**	**80%**	**90%**
P1	75.2%	77.6%	79.1%	79.3%	81%
P2	68.8%	70.8%	72.3%	71.7%	73.3%
P3	69.8%	70.2%	71.6%	73.4%	73.3%
P4	67.8%	71.2%	71.5%	73.3%	73.3%
P5	70.3%	74.3%	75.4%	78.2%	77.9%

4.1.7. Summary of Analysis Based on Culture

The results of six classes keystroke data analysis based on culture are summarized in Table 11 below. Overall, it can be concluded that the Malays VS Chinese had the highest average accuracy rate of 86.02% for the 50% learning ratio. The rate of accuracy increased for every additional 10% learning ratio. This proves that typing for the same user group can be distinguished by category. The more learning data supplied to the system for identification, the higher the accuracy obtained.

Table 11. Summary of average percentages of five sentences from each category by the culture for learning ratio between 50% and 90%

Classes	**Average Learning Ratio**				
	50 %	**60 %**	**70 %**	**80 %**	**90 %**
Malays VS Chinese	86.02%	86.94%	87.66%	87.98%	88.52%
Malays VS Indians	85.16%	85.88%	86.46%	86.70%	87.28%
Chinese VS Indians	78.46%	79.88%	80.62%	81.68%	82.16%
Others VS Indians	76.68%	79.04%	81.02%	82.06%	84.48%
Others VS Chinese	68.24%	71.50%	74.34%	76.48%	77.52%
Others VS Malays	70.38%	72.82%	73.98%	75.18%	75.76%

4.2. Result Based on Education Level Using CGPA

This study uses CGPA as a benchmark to differentiate the educational achievement of a person. The CGPA was divided into two sections, 3.0 and above and 3.0 and below. Based on the statistics in Figure 4, data obtained during the data collection process recorded the number of volunteers who received a CGPA of less than 3.0 was 60. Therefore, the data of volunteers who obtained CGPA 3.0 and above were randomly selected for 60 people to enable the balance of data between the two classes during the analysis process.

Figure 12 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the CGPA 3.0 above and below 3.0 using 5 words. The results in Table 12 showed that recognition performance ranged from 78% to 87% at the 50% learning ratio.

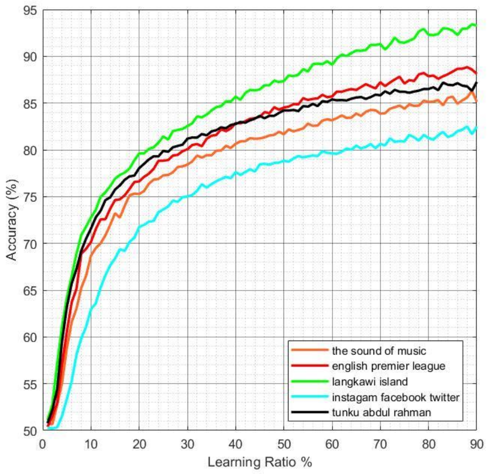

Figure 12. Average accuracy vs Learning Ratio (CGPA > =3.0 VS CGPA <3.0)

Based on the results obtained from the classification using educational level measured using CGPA, it was found that there were good significant differences. This may be due to the number of volunteers involved among university students.

Table 12. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Educational Level - CGPA>= 3.0 VS CGPA < 3.0

Text	Leaning Ratio				
	50%	**60%**	**70%**	**80%**	**90%**
P1	84.21%	85.39%	85.83%	86.50%	87.26%
P2	87.35%	89.09%	91.30	92.30	93.27%
P3	84.52%	85.70%	87.22%	87.86%	88.13%
P4	78.82%	79.62	80.63%	81.26%	82.48%
P5	81.66%	83.16%	83.87%	85.12%	85.14%

4.3. Result Based on Gender

Gender is the final feature of soft biometric used for user classification using KD. Based on the results obtained from Figure 13, it is shown that the classification of gender based on typing methods was between 62% and 80.9% at 50% learning ratio.

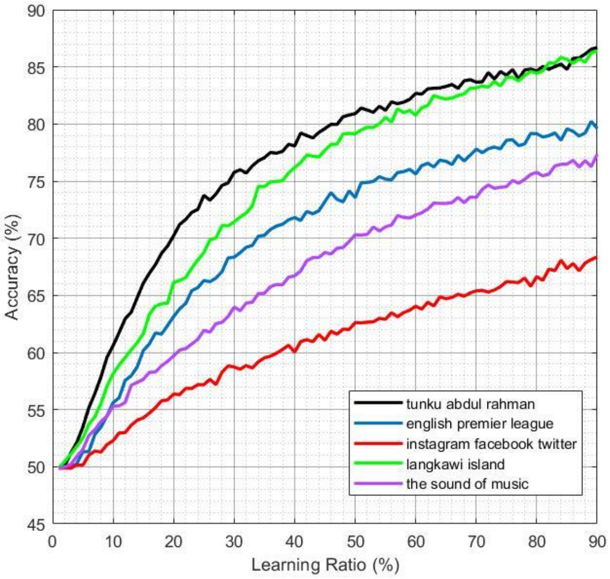

Figure 13. Average accuracy vs Learning Ratio for gender category (Male VS Female)

The results obtained show that the classification of users based on typing style for men and women is indistinguishable because it provides inconsistent results. This may be due to the similar preferences of these two different groups. For example, a man possessed a woman characteristic and vice versa.

4.4. Result Based on Region of Birth

Malaysia consists of 13 states and 3 federal territories. These states can be grouped into 4 sections: North of Peninsular Malaysia, East of Peninsular Malaysia, Central of Peninsular Malaysia, South of Peninsular Malaysia, Sabah and Sarawak [39, 40]. This research focused on the states in Peninsular Malaysia only and each state was grouped according to the breakdown in Table 13.

Table 13. Breakdown of the states in Peninsular Malaysia

Region	States
Northern of peninsular Malaysia	Perlis, Kedah, Pulau Pinang, Perak
Eastern of peninsular Malaysia	Kelantan, Terengganu dan Pahang
Central of peninsular Malaysia	Selangor, Kuala Lumpur, Putrajaya
Southern of peninsular Malaysia	Negeri Sembilan, Malacca, Johor

Based on the overall data obtained from Figure 3 above, the statistical breakdown of data by each region is, Northern (NN) -136; Eastern (EN)-43, Southern (SN) -33 and Central (CL) -33. 5 data are classified as 'Others (OS)' because they are not included in the list of regions of birth studied. The five data are volunteers from Saudi Arabia, Thailand and three Indonesian. The Others category will not be analyzed because the data set obtained is too small. From the statistics, keystroke data for the

Central and Southern region of Peninsular Malaysia is the lowest among the other regions. Volunteers in the north region were the largest participants, i.e. 136. For analysis purposes, the number of participants in each region were made equal. For example, only 33 records randomly selected from the Northern and Eastern regions of Peninsular Malaysia were used to be compared against the Central region. This is to balance the amount of data to be analyzed using SVM. The same is done for other regions of birth by comparing the lowest number of records between the two classes. The total number of records analyzed for the class of each word is shown in Table 14:

Table 14. Breakdown of keystroke records analyzed by region of birth classes - CGPA>= 3.0 VS CGPA < 3.0

Class Category	Total typing	Total no of user for Each Class	Total Keystroke for Each Class
NN vs SN	7	33	231
NN vs CL	7	33	231
NN vs EN	7	43	301
CL vs EN	7	33	231
SN vs EN	7	33	231
CL vs SN	7	33	231

4.4.1. Northern vs. Southern Region

Figure 14 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the northern and southern regions using 5 words. The results obtained were poor because of the 50% learning ratio, the accuracy obtained only reached between 69% to 74% for 4 words, while only 1 sentence reaches 75% which is Langkawi island. This means the typing pattern for volunteers in the Northern and Southern regions cannot be clearly distinguished for these two categories. Table 15 shows a summary of the accuracy obtained for 50% of the learning ratio.

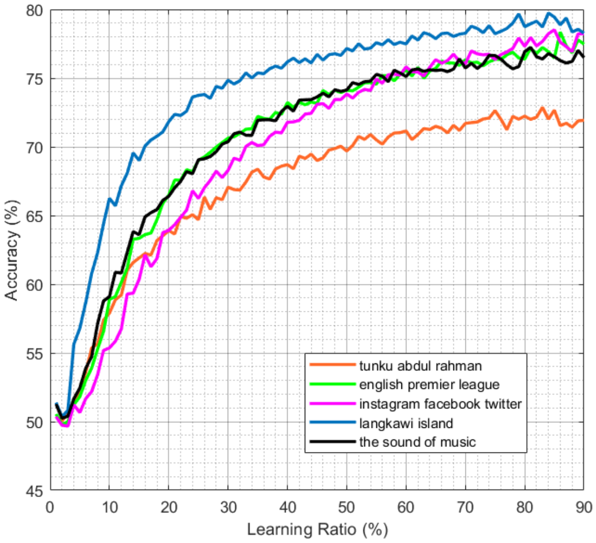

Figure 14. Average accuracy vs Learning Ratio (Northern VS Southern)

Table 15. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Northern VS Southern

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	69.69%	71.15%	71.70%	72.23%	71.94%
P2	77.15%	77.48%	78.26%	78.74%	78.28%
P3	74.13%	75.59%	75.73%	76.91%	77.46%
P4	73.84%	75.81%	76.04%	77.31%	78.27%
P5	74.13%	75.11%	75.73%	76.91%	76.48%

4.4.2. Central vs. Eastern Region

Figure 15 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the central and eastern regions using 5 words. The results obtained were quite good because of the 50% learning ratio, since the accuracy obtained from three sentences exceeded 80%, only two sentences reached between 73% - 78% which is langkawi island

and tunku abdul rahman. Table 16 shows a summary of the accuracy obtained for 50% of the learning ratio.

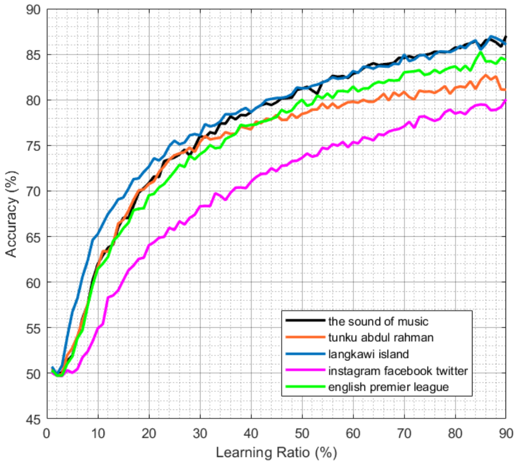

Figure 15. Average accuracy vs Learning Ratio (Central VS Eastern)

Table 16. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Central VS Eastern

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	78.45%	79.78%	80.87%	81.32%	81.10%
P2	81.26%	83.15%	84.97%	85.73%	86.04%
P3	80.01%	81.45%	83.00%	83.66%	84.36%
P4	73.63%	75.36%	77.08%	78.49%	80.00%
P5	81.26%	83.15%	84.97%	85.72%	86.05%

4.4.3. Eastern vs. Southern Region

Figure 16 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the eastern and southern regions using 5 words. The results obtained between the two classes were 75.02% up to 87.44% accuracy for the learning ratio of 50% and above. Table 17 shows a summary of the accuracy obtained for 50% of the learning ratio.

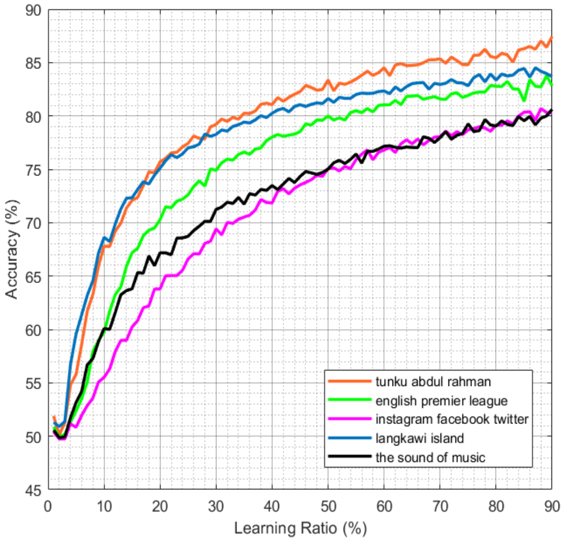

Figure 16. Average accuracy vs Learning Ratio (Eastern VS Southern)

Table 17. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Eastern VS Southern

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	83.37%	84.50%	85.34%	85.44%	87.44%
P2	81.63%	82.33%	82.95%	83.35%	83.96%
P3	79.99%	81.03%	81.58%	82.78%	82.83%
P4	75.02%	76.80%	77.96%	79.05%	80.62%
P5	75.02%	77.15%	77.96%	79.05%	80.63%

4.4.4. Northern vs. Central Region

Figure 17 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the northern and central regions using 5 words. The results obtained between the two classes were 85% up to 93% accuracy for the learning ratio of 50% and above. Table 18 shows a summary of the accuracy obtained for 50% -90% of the learning ratio.

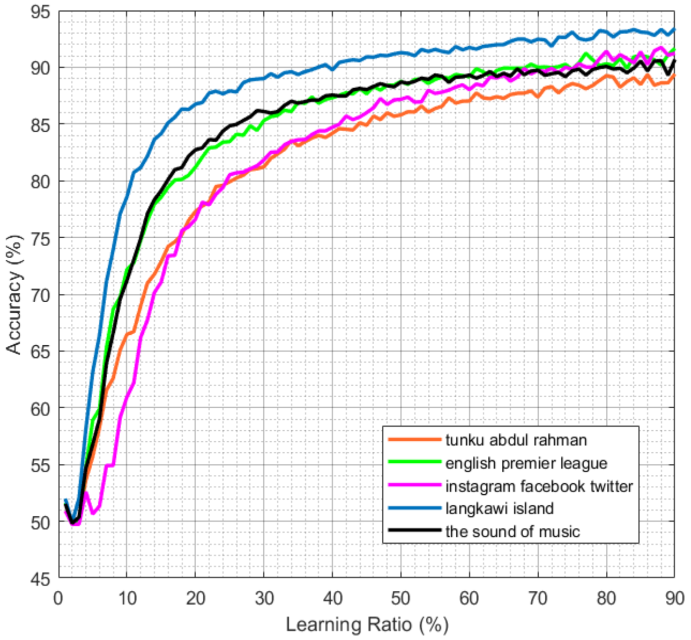

Figure 17. Average accuracy vs Learning Ratio (Northern VS Central)

Table 18. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Northern VS Central

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	85.80%	87.02%	87.36%	89.12%	89.40%
P2	91.28%	91.74%	92.46%	93.11%	93.42%
P3	88.69%	89.30%	89.73%	90%	91.66%
P4	87.17%	88.03%	89.73%	91.39%	90.68%
P5	88.69%	89.30%	89.73%	90.06	90.68%

4.4.5. Central vs. Southern Region

Figure 18 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the central and southern regions using 5 words. The results obtained between the two classes were 70.65%% up to 83.21%% accuracy for the learning ratio of 50% and above. Table 19 shows a summary of the accuracy obtained for 50% -90% of the learning ratio.

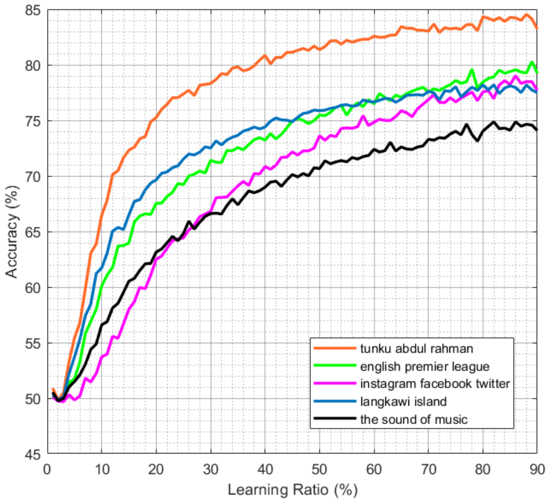

Figure 18. Average accuracy vs Learning Ratio (Central VS Southern)

Table 19. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Central VS Southern

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	81.34%	82.56%	83.02%	84.31%	83.21%
P2	75.86%	76.64%	77.59%	78.19%	77.48%
P3	75.47%	76.64%	77.56%	78.19%	79.58%
P4	73.61%	74.82%	76.85%	77.54%	77.72%
P5	70.65%	72.39%	73.28%	74.03%	74.56%

4.4.6. Northern vs. Eastern Region

Figure 19 shows the results obtained for an average of 100 times the iterations of recognition rate accuracy with the learning ratio between the northern and eastern regions using 5 words. Table 20 shows the results obtained between the two classes can be categorized as good because three of the five sentences tested had more than 80% accuracy, while the remaining two sentences had 77.07% and 79.71% accuracy.

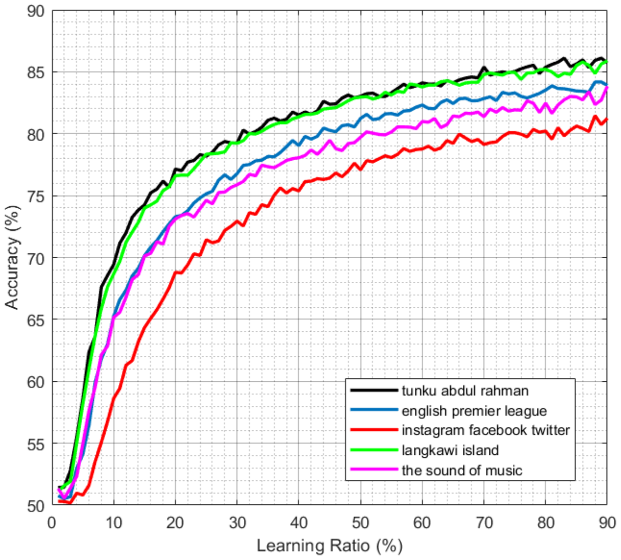

Figure 19. Average accuracy vs Learning Ratio (Northern VS Eastern)

Table 20. Summary of the accuracy obtained from 50% to 90% of the learning ratio for Northern VS Eastern

Text	Leaning Ratio				
	50%	60%	70%	80%	90%
P1	83.02%	84.12%	85.39%	85.38%	85.72%
P2	82.95%	83.77%	84.78%	85.19%	85.27%
P3	81.22%	82.31%	82.84%	83.52%	83.90%
P4	77.07%	78.76%	79.13%	80.21%	81.24%
P5	79.71%	80.98%	81.37%	82.47%	83.83%

4.4.7. Summary Analysis Based on Region of Birth

The results of six classes keystroke data analysis based on region of birth can be summarized as illustrated in Table 21. Overall, typing classification using region of birth gave a rather impressive result because at the 50% learning ratio, the lowest accuracy rate was 73.79%%, while the highest accuracy was 91.17%. The most distinguishable typing method was for the North versus Central and North versus East categories due to at 50% SVM learning, the accuracy obtained was over 80%.

Table 21. Average percentages of five sentences from each category by the region of birth for learning ratio between 50% and 90%

Classes	Average Learning Ratio At				
	50%	60%	70%	80%	90%
NN VS SN	73.79	75.03	75.49	76.42	76.49
CL VS EN	78.92	80.58	82.18	82.98	83.51
EN VS SN	79.01	80.36	81.16	81.93	83.10
NN VS CL	88.33	89.08	89.80	89.92	91.17
CL VS SN	75.39	76.61	77.66	78.45	78.51
NN VS EN	80.79	81.99	82.70	83.35	83.99

5. Summary

It can be concluded from the results obtained that the application of the soft biometric elements in the Keystroke Dynamic study can be used for several categories. This classification proves that the combination of soft biometric and KD can be used as an additional security feature in system authentication. The system can compare user profiles detected by using the keyboard and profile registered in the system. The results are expected to help other researchers study the different aspects of soft biometric that can be used to classify users by typing. The results show that the best classifications of typing pattern that can be identified via soft biometric are culture and region of birth. The best classification for culture is obtained from Malay vs. Chinese category with an average reading accuracy of 88.52% at 90% learning ratio. The best classification for region of birth was obtained from the Northern vs. Central category with an average reading accuracy of 91.17% at 90% learning ratio. This shows that there are clear differences in typing patterns for a culture and region of birth. This may be because the writing and speaking languages used by each culture are quite different.

The results of this study can be used in daily environment, especially in the field of computer forensic. With the existence of a database to record the typing patterns of each group of people, KD can help the authorities in identifying groups of cyber-criminals who commit offenses. In addition, this KD can be used in the control of access to all systems that use a username and password where it can be used as a second security filter after the username and password. To further enhance the study in the field of KD and soft biometric, future researchers can use other soft biometric elements in the study of KD and use different identification techniques such as Fuzzy Logic and Neural Network.

References

[1] Monrose, F. and A.D. Rubin, *Keystroke dynamics as a biometric for authentication.* Future Generation Computer Systems, 2000. 16(4): p. 351-359.

[2] Montalvão, J., et al., *Contributions to empirical analysis of keystroke dynamics in passwords.* Elsevier International Journal of Pattern Recognition Letters, 2015. 52: p. 80-86.

[3] Idrus, S.Z.S., et al. *A preliminary study of a new soft biometric finger recognition for keystroke dynamics*. 2012.

[4] Ailisto, H., et al., *Soft biometrics—combining body weight and fat measurements with fingerprint biometrics.* Elsevier International Journal of Pattern Recognition Letters, 2006. 27(5): p. 325-334.

[5] Arigbabu, O.A., et al., *Integration of multiple soft biometrics for human identification.* Elsevier International Journal of Pattern Recognition Letters, 2015. 68: p. 278-287.

[6] GALTON, F., *Signaletic Instructions, including the Theory and Practice of Anthropometrical Identification*. 1896.

[7] Heckathorn, D.D., R.S. Broadhead, and B. Sergeyev, *A methodology for reducing respondent duplication and impersonation in samples of hidden populations.* Journal of Drug Issues, 2001. 31(2): p. 543-564.

[8] Nalla, P.R. and A. Kumar, *Toward More Accurate Iris Recognition Using Cross-Spectral Matching.* IEEE Transactions on Image Processing, 2017. 26(1): p. 208-221.

[9] Park, U. and A.K. Jain, *Face matching and retrieval using soft biometrics.* IEEE Transactions on Information Forensics and Security, 2010. 5(3): p. 406-415.

[10] Wayman, J. *Large-scale civilian biometric systems-issues and feasibility*. in *Proceedings of Card Tech/Secur Tech ID*. 1997.

[11] Idrus, S.Z.S., et al. *Soft biometrics database: A benchmark for keystroke dynamics biometric systems*. in *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*. 2013. IEEE.

[12] Jain, A.K., S.C. Dass, and K. Nandakumar, *Soft biometric traits for personal recognition systems*, in *Biometric Authentication*. 2004, Springer. p. 731-738.

[13] Marcialis, G.L., F. Roli, and D. Muntoni, *Group-specific face verification using soft biometrics.* Journal of Visual Languages & Computing, 2009. 20(2): p. 101-109.

[14] Ambalakat, P. *Security of biometric authentication systems*. in *proceedings of 21st Computer Science Seminar*. 2005.

[15] Ngo, T.T., et al., *Similar gait action recognition using an inertial sensor.* Elsevier International Journal of Pattern Recognition, 2015. 48(4): p. 1289-1301.

[16] Ran, Y., G. Rosenbush, and Q. Zheng. *Computational approaches for real-time extraction of soft biometrics*. in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. 2008. IEEE.

[17] Tiwari, S., A. Singh, and S.K. Singh, *Fusion of ear and soft-biometrics for recognition of newborn.* Signal and Image Processing: An International Journal, 2012. 3(3): p. 103-116.

[18] Zhang, W., et al. *A hybrid emotion recognition on android smart phones*. in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. 2013. IEEE.

[19] Yang, L., et al., *Exploring soft biometric trait with finger vein recognition.* Neurocomputing, 2014. 135: p. 218-228.

[20] Best-Rowden, L. and A.K. Jain, *Longitudinal study of automatic face recognition.* IEEE transactions on pattern analysis and machine intelligence, 2017.

[21] Epp, C., M. Lippold, and R.L. Mandryk. *Identifying emotional states using keystroke dynamics*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011. ACM.

[22] Fairhurst, M. and M. Da Costa-Abreu. *Using keystroke dynamics for gender identification in social network environment*. in *Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on*. 2011. IET.

[23] Giot, R. and C. Rosenberger, *A new soft biometric approach for keystroke dynamics based on gender recognition.* International Journal of Information Technology and Management, 2012. 11(1-2): p. 35-49.

[24] Gunawardhane, S.D., et al. *Non invasive human stress detection using key stroke dynamics and pattern variations*. in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. 2013. IEEE.

[25] Nahin, A.N.H., et al., *Identifying emotion by keystroke dynamics and text pattern analysis.* Behaviour & Information Technology, 2014. 33(9): p. 987-996.

[26] Bakhtiyari, K., M. Taghavi, and H. Husain. *Implementation of emotional-aware computer systems using typical input devices*. in *Asian Conference on Intelligent Information and Database Systems*. 2014. Springer.

[27] Idrus, S.Z.S., et al., *Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords.* Elsevier International Journal of Computers & Security, 2014. 45: p. 147-155.

[28] Idrus, S.Z.S., et al. *Soft biometrics for keystroke dynamics*. in *International Conference Image Analysis and Recognition*. 2013. Springer.

[29] Idrus, S.Z.S., *Soft Biometrics for Keystroke Dynamics.(Biométrie douce pour la dynamique de frappe au clavier)*. 2014, University of Caen Normandy, France.

[30] Idrus, S.Z.S., et al. *Keystroke dynamics performance enhancement with soft biometrics*. in *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*. 2015. IEEE.

[31] Antal, M. and G. Nemes. *Gender recognition from mobile biometric data*. in *Applied Computational Intelligence and Informatics (SACI), 2016 IEEE 11th International Symposium on*. 2016. IEEE.

[32] Idrus, S.S., E. Cherrier, and C. Rosenberger. *Fusion et biométrie douce pour la dynamique de frappe au clavier*. 2016.

[33] Kołakowska, A., *Usefulness of Keystroke Dynamics Features in User Authentication and Emotion Recognition*, in *Human-Computer Systems Interaction: Backgrounds and Applications 4*, Z.S. Hippe, J.L. Kulikowski, and T. Mroczek, Editors. 2018, Springer International Publishing: Cham. p. 42-52.

[34] Katerina, T. and P. Nicolaos, *Mouse behavioral patterns and keystroke dynamics in End-User Development: What can they tell us about users' behavioral attributes?* Computers in Human Behavior, 2018. 83: p. 288-305.

[35] Idrus, S.Z.S., et al. *Keystroke Dynamics for Construction Industry: A Review on Biometric Systems*. in *Applied Mechanics and Materials*. 2015. Trans Tech Publ.

[36] Seah, L.H., et al., *STR Data for the AmpFlSTR Identifiler loci in three ethnic groups (Malay, Chinese, Indian) of the Malaysian population.* 2003. 138(1-3): p. 134-137.

[37] Chen, G., W.J.I. Xie, and V. Computing, *Pattern recognition with SVM and dual-tree complex wavelets.* 2007. 25(6): p. 960-966.

[38] Chen, S., S. Billings, and P.J.I.J.o.C. Grant, *Recursive hybrid algorithm for non-linear system identification using radial basis function networks.* 1992. 55(5): p. 1051-1070.

[39] Abdurofi, I., et al., *Economic analysis of broiler production in Peninsular Malaysia.* 2017. 24(2): p. 761.

[40] Gomez, E.T., *State of Malaysia*. 2004: Routledge.