International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Dec-2022)

https://dx.doi.org/10.12785/ijcds/1201119

Forensic Fingerprint Analysis using Self-Organizing Maps, Classification and Regression Trees and Naïve Bayes Methods

Loong Chuen Lee[1], Nur Izzati Bohari[2], Siti Norfaraan Abd Sanih[1] and Muhammad Yusuf Adam[1]

[1] Forensic Science Program, CODTIS, FSK, Universiti Kebangsaan Malaysia, Bangi, Malaysia
[2]Institute of IR4.0,Universiti Kebangsaan Malaysia, Bangi, Malaysia

Received 19 Mar. 2021, Revised 17 Jul. 2022, Accepted 09 Aug. 2022, Published 31 December 2022

Abstract: Fingerprint is one type of physical evidence frequently encountered at a crime scene. It is useful in revealing identity of the culprit. However, poor quality latent fingerprint collected from a crime scene seldom makes an identification reliable. In practice, the identification is accomplished by matching a known print to an unknown print according to the types and locations of their minutiae features. When it is infeasible to conduct an identification, forensic scientist can attempt to predict the sex of donor of the latent fingerprint in order to narrow down the scope of searching of the suspect. In the context of forensic science, sexual dimorphism in ridge count has been studied for a few decades ago. Meanwhile, gender classification based on fingerprint images have been regularly reported in the field of computer science. Viewed from a practical perspective, image of a latent print collected from a real crime scene can be low in quality or even incomplete. Hence, this study has not considered image of fingerprint as input data but the number of diagonal ridge counted manually within a well-defined region*,* *i.e.* 25 centimeter squared. Firstly, the fingerprint data was explored using self-organizing maps method with respect to sexual dimorphism and ethnic difference. Next, Naïve Bayes (NB) and Classification and Regression Trees (CART) algorithms were, respectively, used to construct predictive model for discriminating gender based on the fingerprint data. A multitude of prediction models were constructed by considering ten-digit, five-digit and one-digit samples, respectively, to predict gender by three races, *i.e.* Chinese, Indians and Malays; and the combined sub-population. Each of the models was validated using bootstrapping without replacement approach. Results showed that the single-digit samples produced accuracy rate slightly lower than that obtained using five- or ten-digit samples. Comparing to the global predictive model, ethnicity-specific models of Indian and Malay subjects showed slight improvement in external accuracy rate. Moreover, by considering all five digits of a particular hand as input data, NB tends to outperform CART. However, both NB and CART are comparable to each other when one-digit sample was considered as input data. In conclusion, SOM is useful for exploring sexual dimorphism of fingerprint data and NB outperforms CART in modelling of the fingerprint data.

Keywords: Naïve Bayes, CART, self-organizing maps, gender, ridge count, Malaysia

1. INTRODUCTION

Fingerprint refers to the pattern seen on the finger of mammals of which formed by the unique configuration of ridges. The fingerprint pattern is known to be a product of genetic and environmental factors. For these reasons, even monozygotic twins never show exactly alike fingerprint pattern. In other words, fingerprint can be used to identify an individual.

Owing to the individuality of fingerprint, it is the preferred evidence adopted by the forensic analyst to identify the potential suspect. Besides that, fingerprint is also frequently employed by government agencies as a biometric marker [1]. Specifically, identification of culprit based on the fingerprint leave at the scene of crime is one crucial task in forensic investigation, [2].

Studies of forensic application of fingerprint can be very diverse. Inherently, due to the presence of dead skin in the deposit of fingerprint, researcher has attempted extracting DNA from the latent fingerprint [3]. Another division of forensic fingerprint analysis is dealing with latent fingerprint detection techniques [4]. However, it is generally agreed that fingerprint identification procedure is the primary concern in forensic fingerprint analysis [5].

E-mail address: lc_lee@ukm.edu.my

In most of the times, latent fingerprint found at a crime scene is incomplete or of low quality and thus rarely sufficient for an accurate identification. Under such circumstances, determining gender of the donor of the questioned fingerprint is still relevant. By knowing the gender of the latent fingerprint, workload of the investigative officer could be lighten by focusing on a particular gender in searching the potential suspect.

This following of the paper composes of five sections. Section 2 discusses works that employed machine learning in forensic fingerprint analysis. Implementation of machine learning methods proposed in this work are explained in section 3. Section 4 provides the results and discussion. Next, section 5 presents a brief conclusion and contribution of the work. Last but not least, limitations and future scopes of the work are addressed in section 6.

2. RELATED WORKS

Papers reporting application of machine learning methods in fingerprint analysis are abundant in the literature, specifically in the domain of computer science. Classification and gender discrimination of fingerprint are two of the most studied topics. Awad [6] and Noor et al. [7] respectively presented a survey of machine learning algorithms applicable to fingerprint classification. Meanwhile, brief reviews on gender discrimination based on fingerprint and machine learning were also reported in the recent literature [8, 9]

Strictly speaking, sex discrimination is much relevant than the fingerprint classification in the context of forensic investigation. In particular, sexual dimorphism in diagonal ridge count of fingerprint was first published by [10]. After the study, a multitude of similar works but employing different populations have been reported in the literature [*e.g.* 11-15]. However, all these studies have not employed any machine learning algorithm to construct a prediction model and have relied on hand magnifying lens to calculate the number of ridge.

On the contrary, researchers from the domain of computer science actively explored various machine learning algorithms in discriminating gender based on multiple features of fingerprint [16-25]. Compared to [11-15] performing manual calculation of number of ridge, [16-25] tended to extract fingerprint feature assisted by either image processing technique or digitizing the image of fingerprint. For instance, [20] proposed an image processing algorithm to determine the number of ridge. Despite that, an improper processing might introduce artifacts that rendering data inaccuracy.

As this work devoted to forensic investigation purpose, we opted to follow procedure proposed by [10], *i.e.* manually determine number of ridge within a confined region on fingerprint. However, to reduce human error during calculation, a hand-held microscope instead of a hand magnifying lens was employed to assist in calculating the number of ridge. Then, the calculation was performed on the zoomed image.

To the best of our knowledge, only several work have assessed performance of machine learning using ridge counts data of fingerprint [21-25]. The rest of the works have considered features that can hardly be extracted from a low-quality image of fingerprint, *e.g.* minutiae [17]. Table 1 summarizes details of the works [14, 15, 20-25]. Studies that excluded diagonal ridge count as one of the fingerprint feature were not listed in Table I. Additionally, two studies considered Malaysian subjects but have not employed any machine learning method were also listed in Table I [14, 15].

To the best of our knowledge, there are two studies dedicated to Malaysian population [14, 15]. It is important to note that Malaysia's population is a multiracial population. This is attributed to the British colonialism happened in the second half of the 19th century. However, Nayak et al. [14] and Abdullah et al. [15] have not considered the ethnic difference; and the number of subject was also quite small, *i.e.* 100 [14] and 50 [15] Malaysians. Most importantly, the authors have not utilized any machine learning methods in assessing the sexual difference but only mean and standard deviation values.

It is worth mentioning that the input data employed by [21-25] were the digital images of fingerprint. Consequently, the researchers have extracted more than one fingerprint features, including ridge count, for constructing model to predict sex of donor. Eventually, it is observed that all the machine learning methods listed in Table I have achieved accuracy rate above 80%. In particular, [25] found that Naïve Bayes (NB) classifier was outperformed decision tree. Generally, it is noted that performance of machine learning when the number of fingerprint and feature are increased.

In order to evaluate the ethnic difference, we have included 100 Chinese, 100 Indians and 100 Malays with equal proportion in male and female. This is the first work attempted on evaluating ethnic difference among Malaysian based ridge count of fingerprint using machine learning method. The sexual dimorphism of the fingerprint data has been demonstrated using descriptive statistics [26-28]. Moreover, it is also of particular interest to evaluate relative performance of NB and decision tree (i.e. Classification and Regression Tree (CART)) by using rather simple fingerprint data, *i.e.* only ridge count of fingerprint of ten digits.

Additionally, it is important to stress that none of the work listed in Table I, except [14] explicitly mentioning that they have studied all the ten digits of an individual. As such, this work appears to be the first comprehensive study evaluating all the ten digits of 300 Malaysians for studying sexual dimorphism and ethnic difference by using machine learning methods.

TABLE I. Works studying sexual dimorphism using fingerprint ridge counts

Authors (Year)	# Subjects (Male: Female)	Fingerprint features (#Digit)	Classifier (Accuracy)
Nayak et al. (2010) [14]	100 Malaysians (50:50)	Diagonal ridge counts (All ten digits)	Not applicable
Abdullah et al. (2015) [15]	50 Malaysians (25:25)	Diagonal ridge counts (Thumb)	Not applicable
Abdullah et al. (2016a) [20]	3000 (1430: 1570)	Diagonal ridge counts (NA)	Not applicable
Abdullah et al. (2016b) [21]	296 Malaysian	Ridge count, Ridge density, RIdge thickness to valley thickness ratio, white lines count (NA)	Decision tree (J48) (96.28%)
Abdullah et al. (2016c) [22]	3000 (1430: 1570)	Ridge count, ridge density, ridge thickness to valley thickness ratio, white line count (NA)	SVM (96.95%) MPLNN Baye net KNN
Abdullah et al. (2016d) [23]	3000 (1430: 1570)	ridge density, ridge thickness to valley thickness ratio, white line count (NA)	MPNN (97.25%)
Alam et al. (2019) [24]	420	White lines count, Ridge thickness valley ratio, Ridge Width, Ridge density (NA)	SVM (88%)
Ceyhan & Sagiroglu (2014) [25]	600 Turkish (300:300)	Ridge thickness, ridge counts, avergae ridge thickness (NA)	Naïve Bayes (94.7– 95.3%) kNN (93.7– 94.0%) DT (94.2– 94.3%) SVM (93.8– 94.2%)

Hence, the purpose of this work is of two-fold: (a) evaluate ethnic difference and sexual dimorphism of ridge count data using self-organizing maps; and (b) construct prediction model using Naïve Bayes (NB) and Classification and regression trees (CART) algorithms, respectively. Performances of the two algorithms were rigorously assessed by varying the proportion of test set, employing one-digit, five-digit and ten-digit respectively as input data; and considering ethnic-specific data.

3. Methodology

A. *Fingerprint data*

Fingerprint data was the collection of ten digits of 300 healthy Malaysian subjects of three major races, *i.e.* Malays, Indians and Chinese. Each of the races consisted of 50 male and 50 female subjects. Figure 1 illustrates the primary steps employed in obtaining and processing the 3000 fingerprints. All the selected subjects full-filled the following characteristics:

1. Have no mixed marriage over three generations.
2. The fingerprint pattern is free from any defects, *e.g.* cuts.
3. Do not suffer from any skin disease.

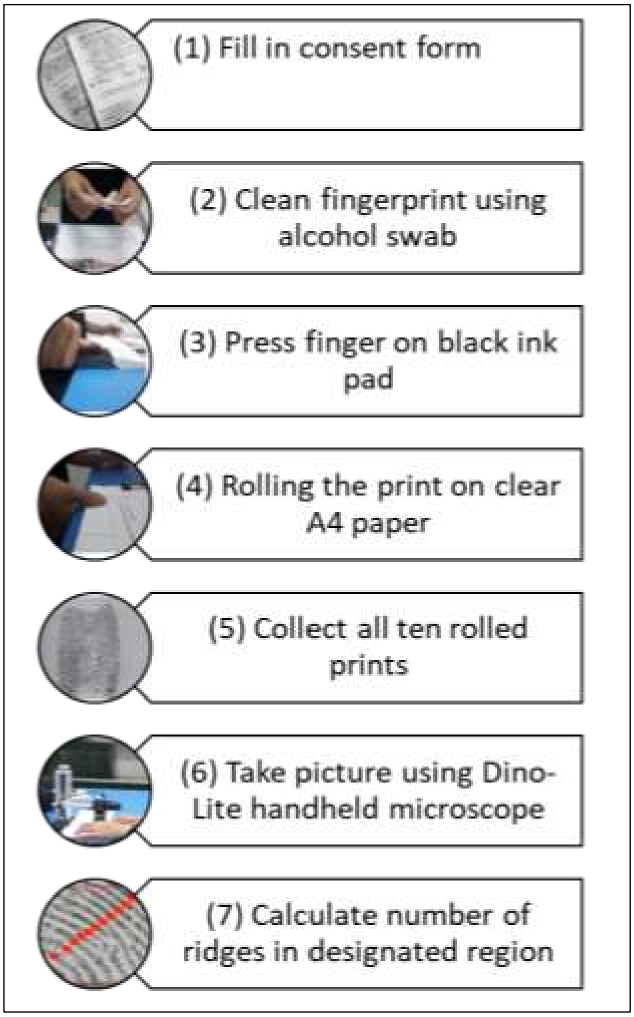

(1) Fill in consent form
(2) Clean fingerprint using alcohol swab
(3) Press finger on black ink pad
(4) Rolling the print on clear A4 paper
(5) Collect all ten rolled prints
(6) Take picture using Dino-Lite handheld microscope
(7) Calculate number of ridges in designated region

Figure 1. Procedures employed in collecting and processing fingerprint sample.

Each subject was first asked to clean their fingertips using alcohol swab. Then, all the ten digits were rolled on the form. A Dino-Lite USB digital microscope was used to capture zoomed view of the fingerprints. Before that, the print was divided into four zones according to the guidelines proposed by Acree [10]. Figure 2 indicates how the designated region was selected by the three primary fingerprint patterns. Then, the number of diagonal ridges were determined within a pre-defined area (5mm x 5mm).

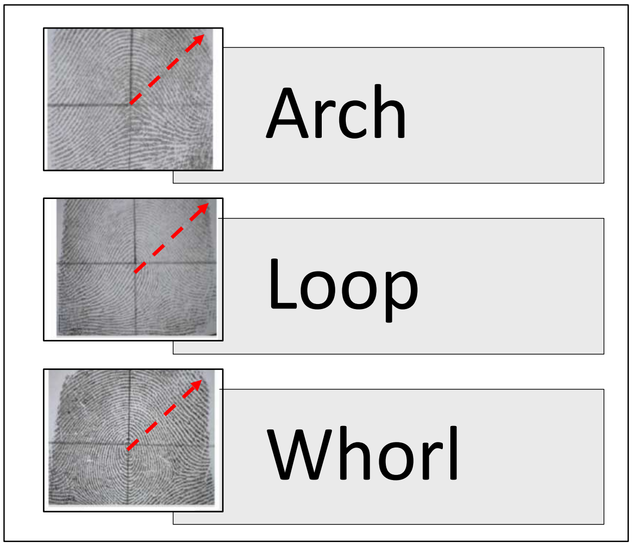


Figure 2. Direction of calculating number of ridges in three main types of fingerprint pattern, *i.e.* arch, loop and whorl

B. Rearrangement of fingerprint data

The principal fingerprint data composed of 300 subjects and number of diagonal ridge across the ten digits. In order to gain more insights into the data, we have rearranged the primary data into 52 different sub-datasets by considering the digits individually, by right and left hands, respectively according to Malaysian and then by the three ethnic groups. Table II summarise the compositions of the 52 sub-dataset.

Firstly, the primary data was split according to ethnicity, *i.e.* Malays, Chinese and Indians. The purpose was to evaluate sexual dimorphism using ethnicity-specific model. The modelling process was recursively performed by using one-digit, five-digit (left or right hand) and ten-digit (both hands) samples in sequential as input data.

Secondly, another 13 sub-datasets were prepared for predicting gender regardless of the ethnicity by using one-digit, five-digit (left or right hand) and ten-digit (both hands) samples in sequential as input data. The aim was to prepare a set of global predictive models that can be used for assessment of merits of ethnic-specific predictive models.

TABLE II. REARRANGEMENT OF THE PRINCIPAL FIGNERPIRNT DATA BY ETHNICITY AND DIGIT

RACE	SEX	THUMB	INDEX	MIDDLE	RING	LITTLE
MALAYS	R	M.R1	M.R2	M.R3	M.R4	M.R5
	L	M.L1	M.L2	M.L3	M.L4	M.L5
CHINESE	R	C.R1	C.R2	C.R3	C.R4	C.R5
	L	C.L1	C.L2	C.L3	C.L4	C.L5
INDIANS	R	I.R1	I.R2	I.R3	I.R4	I.R5
	L	I.L1	I.L2	I.L3	I.L4	I.L5
M'SIAN	R	A.R1	A.R2	A.R3	A.R4	A.R5
	L	A.L1	A.L2	A.L3	A.L4	A.L5

C. Data analysis

The fingerprint data was first explored using self-organizing maps (SOMs) method. The purpose was to evaluate the potential of the ten digits individually and then by hands, in discriminating males and females among the 300 Malaysians and then by the three ethnic groups, respectively.

After that, the data was further modelled using two supervised methods, *i.e.* classification and regression tress (CART) and naïve Bayes (NB) algorithms. Figure 3 explains the general modelling procedures of CART and NB algorithms as well as the respective model evaluation procedures.

The models were rigorously assessed using six series of training and test sets prepared by varying the number of repeats in random resampling without replacement (9 and 99); and split ratio (9:1, 8:2 and 7:3), respectively. Consequently, stability of the predictive models were derived by comparing the six mean external accuracy rates of a given data and modelling method.

The predictive accuracy (Acc) of the CART and NB models are determined as follows:

$$\text{Acc} = \frac{n'}{n} \quad (n' \leq n) \tag{1}$$

where n refers to the total number of training/test samples and n' denotes the number of correctly predicted training/test samples.

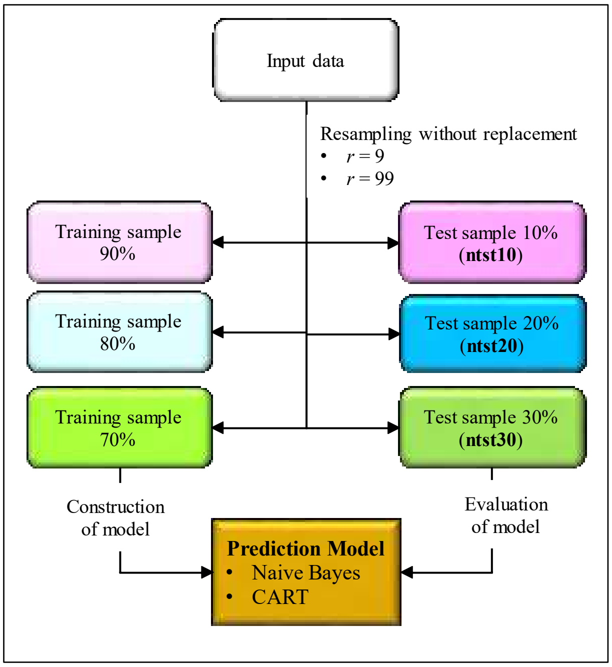


Figure 3. Overview of modelling and validation of prediction models contructed from the fingerprint data

Due to the overwhelming number of models constructed in this work, SOM was also employed herein to examine the relative performances of CART and NB by the type of input data, *i.e.* one-digit, five-digit and ten-digit data.

D. Machine learning methods

All the machine learning algorithms were implemented by the R software version 3.6.2 [29]. The SOM algorithm was conducted using code prepared by Kohonen [30] package. Meanwhile, the e1071 [31] and rpart [32] packages, respectively, provides the naïve Bayes and CART algorithm to construct the prediction models.

1) Self-organizing maps (SOMs)

Self-organizing maps (SOMs) have been reported in fingerprint analysis studies. For instance, [33] employed SOMs to assess fingerprint image quality. Meanwhile, [34] adopted SOMs in an automatic scheme for the identification of fingerprint images. However, in this work, SOM was employed for: (a) exploring the relative potential of digits in discriminating the sex and race of Malaysian according to the number of diagonal ridge count; and (b) comparing relative performances of a multitude of prediction models.

SOMs also known as Kohonen Map is proposed by [35]. It employs unsupervised learning approach to elucidate latent structure of the input data. SOM algorithm aims to arrange the input patterns (*i.e.* fingerprint data in this work) to a topological structure while preserving the relations between the patterns.

SOM modelling involves a 2-layered network composing of *n* processing units. The first layer of units responsible to connect to the signal; and is linked to the second layer also known as competitive layer (2-dimensional). Each of the connections is associated with a weight value of which is updated during training.

First and foremost, the ethnic difference of the 300 Malaysian subjects was explored using SOM. In this case, the first layer composing of the 300 Malaysian subjects (i.e. processing units) was linked to the competitive layer. After careful optimization, the second layer of 3 x 3 units was employed to connect to the first layer. Parameters of SOMs employed for evaluation of ethnic difference are as follows:

- Number of processing units: 300
- Dimension of topology: 3 x 3
- Type of topology: rectangular
- The number of times the complete data set will be presented to the network: 100

Next, the sexual dimorphism of the fingerprint data was studied by the three ethnic groups. Herein, the first layer was constructed by the 100 Malaysian subjects of a particular ethnicity, *i.e.* Chinese, Indians, and Malays. Due to the reduced number of subjects, the second-layer of 2 x 2 units was found to be preferred over that of 3 x 3. The parameters of SOMs used for evaluation of sexual dimorphism of a given ethnic group based on the ridge count data are listed below.

- Number of processing units: 100
- Dimension of topology: 2 x 2
- Type of topology: rectangular
- The number of times the complete data set will be presented to the network: 100

Last but not least, parameters of SOMs for evaluation of classifiers are as shown below. In this case, the 12 classifiers (2 modelling algorithms x 3 split ratio x 3 input data) formed the first layer of units. The weight value connecting first layer to the second layer was governed by the mean accuracy rate of the classifiers. The purpose of the analysis was to identify the most stable and accurate classifier.

- Number of processing units: 12
- Dimension of topology: 2 x 2

- Type of topology: rectangular
- The number of times the complete data set will be presented to the network: 100

2) *Classification and Regression Tree (CART)*

Many algorithms have been proposed to construct a decision tree model; and Classification and regression trees (CART) is one of the most common ones. CART constructs a prediction model by recursively partitioning the data into two classes. In this case, CART keeps splitting the donors of fingerprints into class of male or female according to the number of ridge of the donor.

The splitting stops once the heterogeneity of data reached to the satisfactory level [36]. Herein, the Gini coefficient was employed for that purpose. Given a binary data, the Gini coefficient can be expressed as below:

$$Gini(T) = 1 - \sum_{g}[P(C_g \mid T)^2] \qquad (2)$$

where T denotes a node in the decision tree model; $P(C_g \mid T)$ refers to the conditional probability of node T being the class of g. In this case, g = 2, *i.e.* male and female. Hence, eq. (3) can be simplified as follow:

$$Gini(T) = 2P(1-P) \qquad (3)$$

3) *Naïve Bayes (NB)*

Naïve Bayes (NB) classifier relies on Bayes rule to build a prediction model [37]. Herein, the prediction problem was formulated as below:

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \qquad (4)$$

where x refers to the fingerprint ridge count; y denotes the sex of the donor. Since the ratio of male and female was 1:1, the $P(x)$ can be ignored in this modelling problem. Meanwhile, $P(y)$ denotes the class *a priori* probabilities. In this work, the male and female are assumed equally likely i.e., $P(y_{\text{male}}) = P(y_{\text{female}})$. Last but not least, $p(x \mid y)$ is calculated using the training data.

NB aims to predict the sex of the donor (y) based on the number of ridge (x). Given an unknown fingerprint with having a particular number of ridge (x), the sex of the donor (y) is predicted as male if the posterior probability of male $p(y_{\text{male}} \mid x)$ is higher than that of female $p(y_{\text{female}} \mid x)$ and otherwise the donor will be assigned to be female.

4. Results and Discussions

A. *Exploring sexual dimorphism using self-organizing maps*

The discriminatory ability of the ten digits was first explored using SOM plot and the respective mapping plot. The former plot shows the relative magnitude of all the studied variables in relation to the samples plotted in the latter plot. Additionally, the mapping plot also presents the distribution of the subjects by a factor of concern, *e.g.* sex and ethnicity.

Figures 4 and 5, respectively presents the sexual and ethnic differences in the 300 Malaysians. Meanwhile, Figures 6 to 8 show the outputs of SOM obtained using fingerprint data by Chinese, Indians and Malays subjects, respectively.

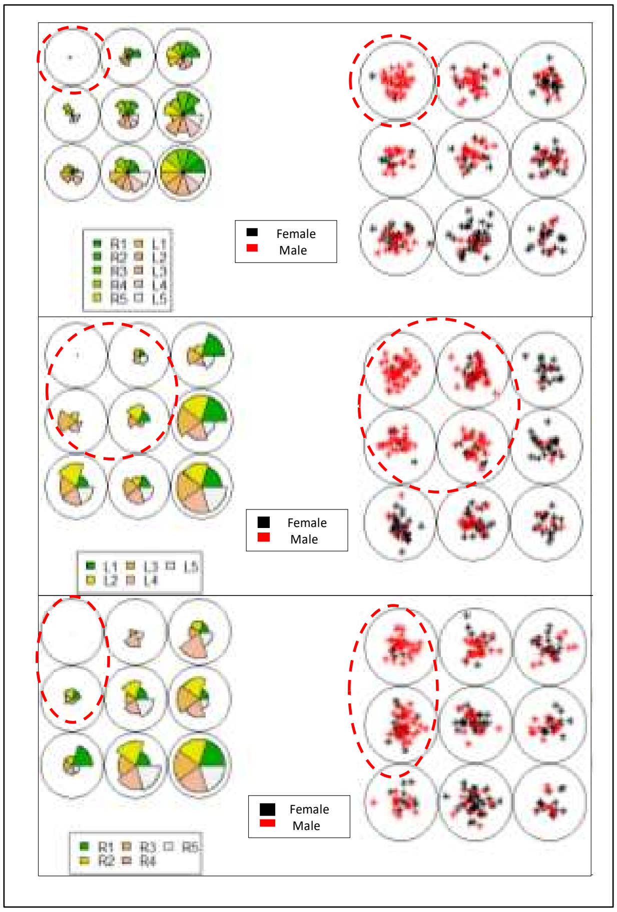


Figure 4. Exploring sexual dimoprhism in 300 Malaysian based on (top) ten-digit; and five-digit of (middle) left and (bottom) right hands, respectively

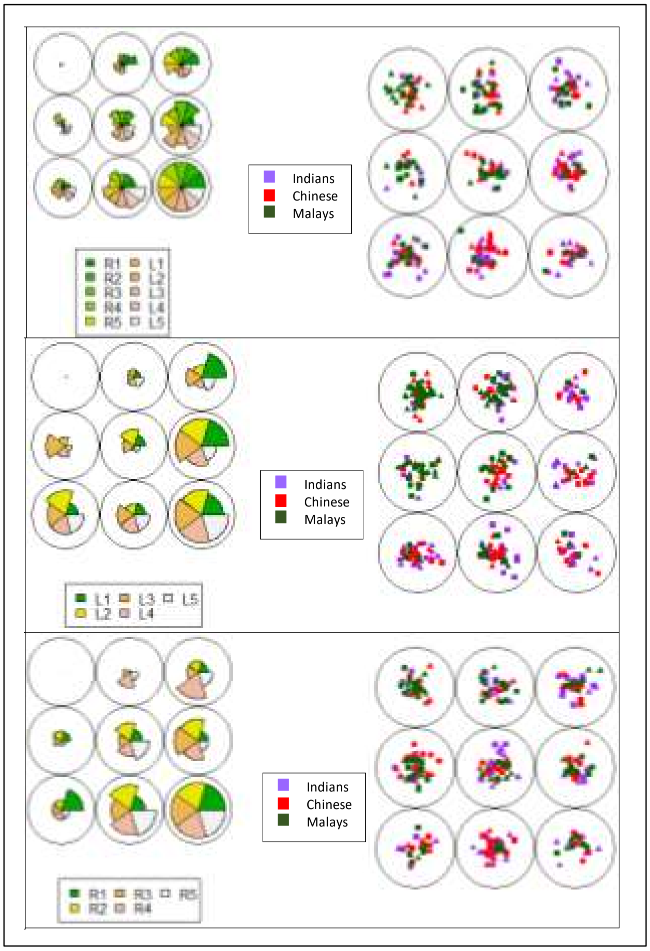


Figure 5. Exploring ethnic variation in 300 Malaysian based on (top) ten-digit; and five-digit of (middle) left and (bottom) right hands, respectively

The number of units in the 2-D topological structure has been carefully optimized. After several trial-and-errors, it was found that 3 x 3 and 4 x 4 were the best options for the 300 Malaysians and 100 ethnicity-specific subjects, respectively. The most optimum topological size is the one that most of the units in the mapping plot are loaded with samples and none of the units is leave empty.

Generally, by referring to Figures 4 and 5, it can be seen that the sexual difference was more significant than the ethnic variation. This is because none of the units (of mapping plot) in Figure 5 appeared to be homogenous, *i.e.* a unit dominated by one colour. On the other hand, several units in Figure 4 were appeared to be homogenous (highlighted by red circle).

The radius of a wedge in the SOM plot (left plots in Figures 4 and 5) indicates the magnitude of the variable in relation to the samples located in the corresponding mapping plot (right plots in Figures 4 and 5).

Generally, most of the females showed significantly high number of ridge count in most of the digits. Meanwhile, males seldom presented high number of ridge count. This is in accordance with that reported by Acree [10]. With respect to the discriminatory ability of the digits, left hand was seen to exhibit higher sexual variations than the respective right hand. Most importantly, sexual dimorphism presented by all the ten fingers has not be better than that observed with the left hands.

In contrast to sexual dimorphism, ethnic variation in the number of ridge count is not significant. Firstly, it is hard to find any units in the mapping plots (right plots in Figure 5) appeared to be homogenous, *i.e.* dominated by one ethnic group. Despite that, in general, it was observed that Malays tended to show low number of ridge count. Meanwhile, Chinese and Indians dominated units exhibiting rather high number of ridge count. Basically, the segregation of the three ethnic groups were not improved even when the input data changed to the five-digit sample. This is different from what was seen in the sexual dimorphism (Figure 4).

Since the sexual dimorphism is apparent than the ethnic difference, we have performed another series of SOM analysis to study the difference between males and females by Chinese, Indians and Malays populations, respectively. Figures 6 to 8 show the SOM outputs obtained from the three different ethnic groups. Because the ten-digit samples has not performed better than the five-digit data, the sexual dimorphism was evaluated only by the left and right hands, respectively.

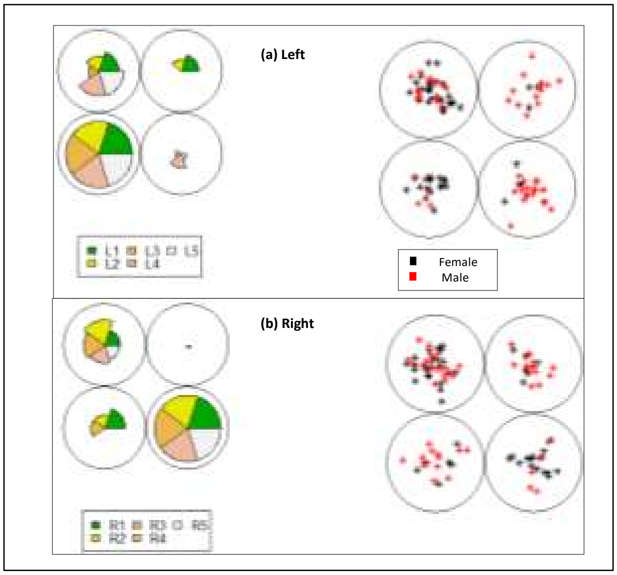


Figure 6. Exploring sexual dimoprhism in 100 Chinese Malaysians based on five-digit of (top) left and (bottom) right hands, respectively

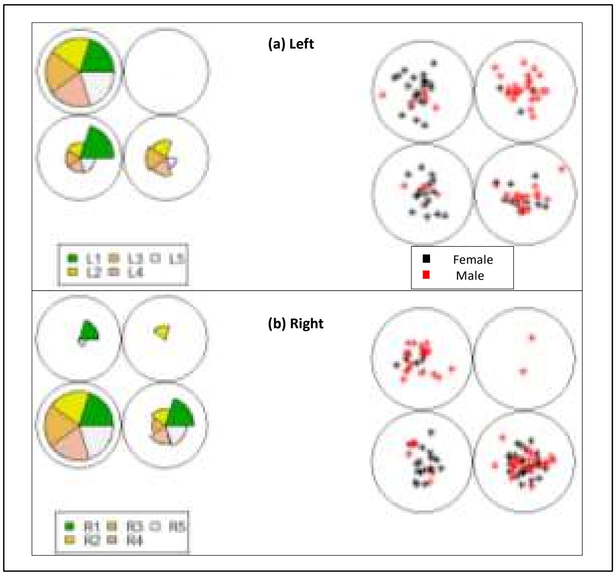


Figure 7. Exploring sexual dimoprhism in 100 Indian Malaysians based on five-digit of (top) left and (bottom) right hands, respectively

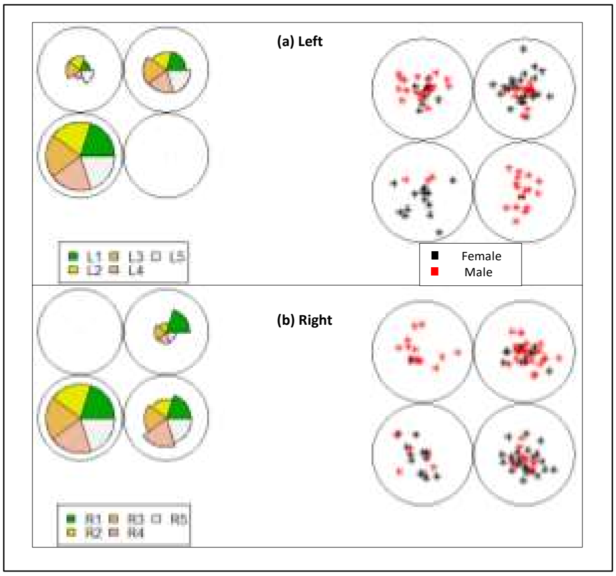


Figure 8. Exploring sexual dimoprhism in 100 Malay Malaysians based on five-digit of (top) left and (bottom) right hands, respectively

Overall, the three ethnic groups showed similar degree of homogeneity in the mapping plots. Firstly, left hand seems to be useful than the corresponding right hand in discriminating the gender. This is because there was more number of homogenous unit when the left hand data was employed in the modelling. Meanwhile, right hand showed unsatisfactory result that all the four units were hardly showed any homogeneity. Additionally, females in all the three ethnic groups were always have more number of ridge count than the males. Hence, this explains why the ethnic variation is much lower than the sexual dimorphism.

B. *Classification performance of NB and CART*

In line with the interpretation obtained *via* SOM technique, NB and CART algorithms were employed to construct model predicting gender and not the ethnicity. As mentioned in section 3, a multitude of predictive models have been constructed by considering varying number of digits as input variables. Figure 9 compares the performances of NB and CART models using one-digit data. On the other hand, Figure 10 shows the relative performances of the models by the five-digit of left and right hands, and ten-digit sample.

By referring to Figure 9, the right little and left thumb fingers were found to be outperformed the other digits regardless of the modelling algorithms. In other words, both the CART and NB models produced similar well performances when right little and left thumb digits were considered as input data.

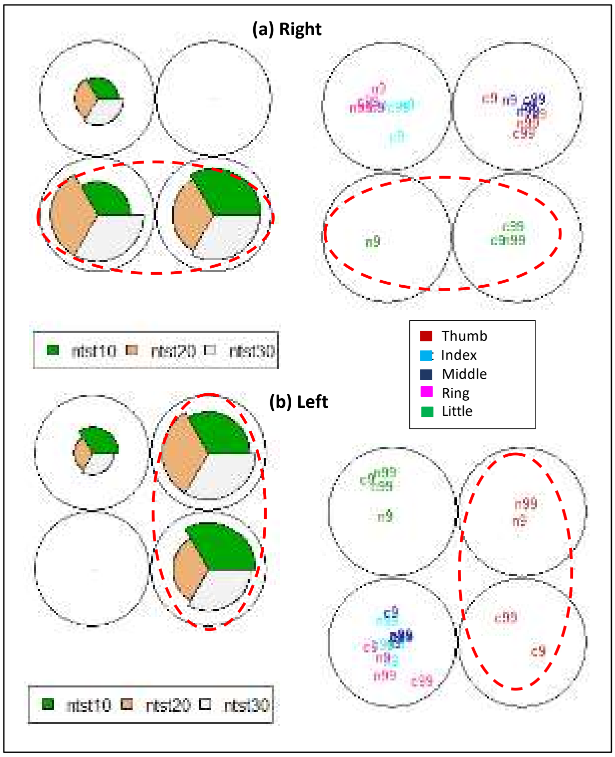


Figure 9. Relative performances of naïve Bayes (NB) and Classification and Regression Trees (CART) models constructed for discriminating sex of 300 Malaysians by using single-digit sample. Each prediction model was validated using three sets of test samples, i.e. 10% (ntst10), 20% (ntst20) and 30% (ntst30) of the data. The mean accuracy rates were estimated from 9 and 99 sets of test samples, respectively.

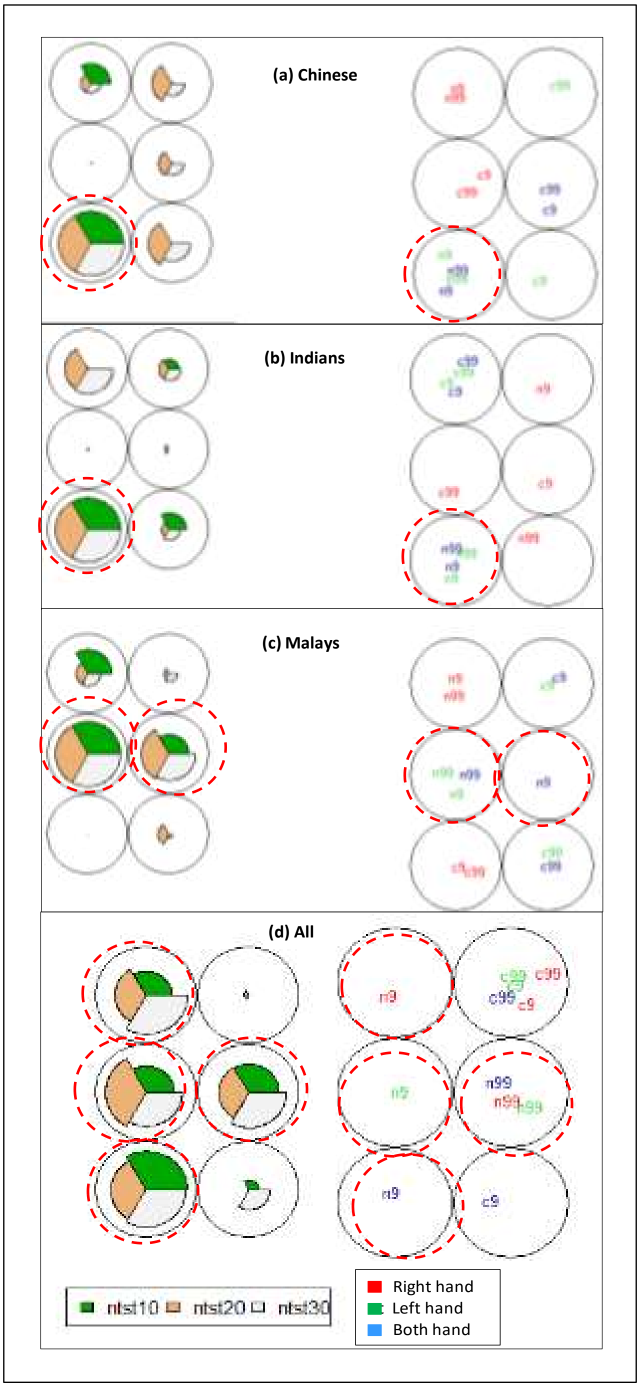


Figure 10. Relative performances of naïve Bayes (NB) and Classification and Regression Trees (CART) models constructed for discriminating sex of (a-c) 100 and (d) 300 Malaysians by using five-digit and ten-digit samples. Each prediction model was validated using three sets of test samples, i.e. 10% (ntst10), 20% (ntst20) and 30% (ntst30) of the data. The mean accuracy rates were estimated from 9 and 99 sets of test samples, respectively.

Based on Figure 10, the NB models were found to outperform the corresponding CART models. By focusing on the ethnicity-specific data (Figure 10 (a-c)), NB models constructed using the five-digits of right hand were never achieved better performance than that built using the five-digit of left hand or even the ten-digit sample. Hence, it seems sound to say that left hand is more reliable than the right hand in predicting sex of the donor.

Additionally, it is worth noting that the mean accuracy rate of NB models estimated from the 9 (n9) or 99 (n99) sets of test samples were much stable than the corresponding CART models, *i.e.* c9 and c99. This is because NB models always achieved similarly high accuracy rates with the nsts 10, ntst20 and ntst30, *i.e.* radius of the three wedges are always maximum. Meanwhile, CART models presented stability similar to the NB models when one-digit sample was used as input data (see Figure 9).

C. *General remaks*

Based on the results, it is obvious that NB models outperformed the CART models. Figure 11 compare the mean external accuracy rates of NB models by input data (single-digit, five-digit and ten-digit samples) and types of subjects (300 Malaysian, 100 ethnicity-specific subjects). Generally, none of the prediction models produced in this study achieved mean accuracy rate approaching 80%.

However, Naïve Bayes was reported by Ceyhan and Sagiroglu [25] to have achieved a predictive accuracy rate above 90%. Similar to this work, the authors also employed Naïve Bayes and decision tree algorithms in constructing predictive model to identify gender based on fingerprint data. Basically, the current work is different from [25] from two aspects.

First and foremost, by combining the three ethnic groups, this work has included 300 subjects. On the contrary, [25] reported the use of 600 subjects, *i.e.* 300 males and 300 females. Hence, the outperformance of NB obtained in this work shall be further verified by including more number of subject.

Despite that, this work and [25] were adopting the similar approach in obtaining the ridge count of fingerprint. Both of the works referred to [3] in selecting the fingerprint region and calculating the ridge counts. However, [8] have considered two extra features, *i.e.* average thickness of fingerprint and total ridge thickness in constructing the prediction model.

Basically, the sexual difference in finger ridge breath has been studied by Mundorff et al. [38]. A total of 5000 fingerprints from 500 subjects (250 males and 250 females) was collected. The sexual difference has been evaluated using descriptive statistics and then ascertained using independent *t*-test test. The authors concluded that females have lower mean ridge breath value than males.

Meanwhile, the topological variation in the number of ridges for gender difference has been evaluated by Gutierrez-Redomero et al. [39]. The fingerprint data was collected from the Spanish population. The number of ridges have been determined from three different region on a finger digit. Despite that, it was noted that the researchers have not employed any machine learning methods in evaluating the fingerprint data.

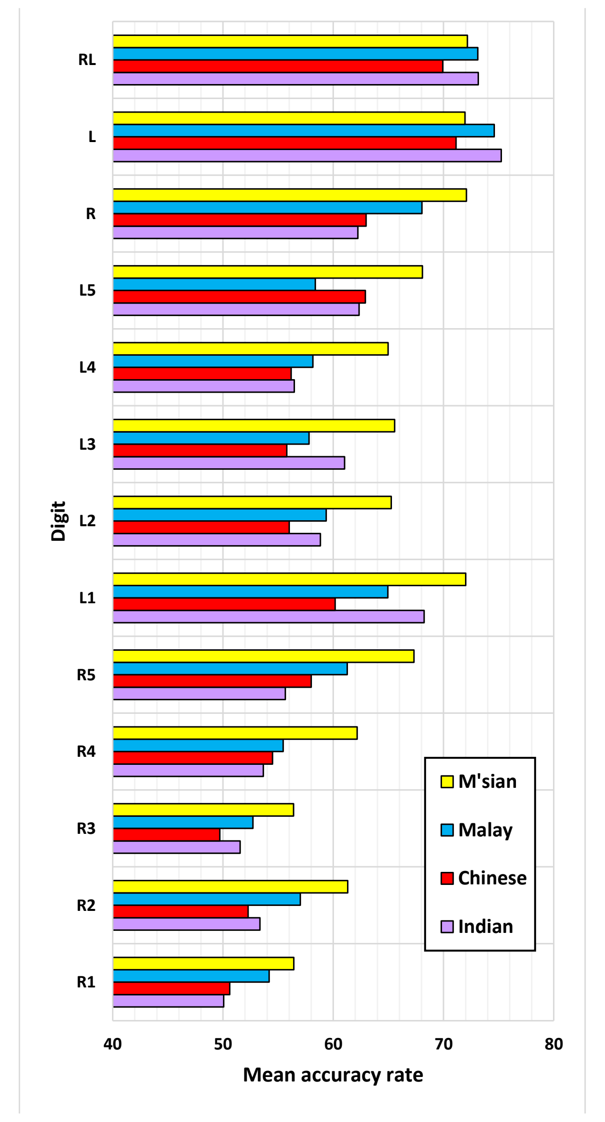


Figure 11. Bar chart showing mean accuracy rates of Naïve Bayes models obtained from 99 sets of test samples (30% of data) by using population of 300 Malaysians, 100 Malays, Chinese and Indians, respecitvely.

Table III listed the best NB prediction models achieved using the ten digits separately, five-digit of left and right hands and ten-digit samples. Basically, the highest external accuracy rate was presented by Indian-subject NB model constructed using five-digit of left hand. On the other hand, the right thumb data produced the lowest accuracy rate by using the global model (*i.e.* 300 Malaysians).

Essentially, four crucial findings derived in this work are: (a) fingerprint ridge count of Indians shows the most apparent sexual dimorphism; (b) digits of left hand can predict gender more accurate than the digits of right hand; (c) left thumb can predict gender as good as the five-digit or ten-digit data; and (d) Naïve Bayes is excellent than CART algorithm in constructing model to predict gender based on fingerprint ridge count.

TABLE III. THE BEST PREDICTIVE NAVIE BAYES MODEL BY INPUT DATA

INPUT DATA	BEST MODEL	EXTERNAL ACCURACY RATE (%)
RIGHT THUMB	MSIAN	**56.41**
RIGHT INDEX	MSIAN	61.32
RIGHT MIDDLE	MSIAN	56.39
RIGHT RING	MSIAN	62.16
RIGHT LITTLE	MSIAN	67.31
RIGHT HAND	MSIAN	72.00
LEFT THUMB	MSIAN	65.26
LEFT INDEX	MSIAN	65.57
LEFT MIDDLE	MSIAN	64.98
LEFT RING	MSIAN	68.09
LEFT LITTLE	MSIAN	72.07
LEFT HAND	INDIAN	**75.24**
BOTH HANDS	INDIAN	73.15

5. CONCLUSION AND CONTRIBUTIONS

This work demonstrated the potential of three popular machine learning methods in forensic fingerprint analysis. Firstly, self-organizing maps is found to be a useful exploratory tool prior to modelling using supervised machine learning methods. Additionally, naïve Bayes is shown to be outperformed Classification and Regression Tree algorithm in predicting sex of fingerprint donor by using ridge count data.

The major contribution of this work is the discussion of three machine learning techniques, *i.e.* Self-organizing maps (SOMs), naïve Bayes (NB) and Classification and Regression Tree (CART) in forensic fingerprint analysis. The findings would encourage more application of machine learning methods in forensic fingerprint analysis, specifically using data obtained in a way applicable to real case investigation.

6. Limitations and Future Work

Despite naïve Bayes is found to be excellent than Classification and Regression Tree in this study. However, the accuracy rate is still need to be improved to a higher level. The performance of the classifiers is somewhat limited by the two facts: (a) rather small number of samples and (b) only one fingerprint feature was being considered, *i.e.* diagonal ridge counts.

In order to address the limitations, the study can be expanded in future by increasing the number of fingerprint donor and to extract more features from a fingerprint. The features that are feasible to be obtained from a low quality print image including diagonal ridge counts from other topological area and the pattern type, *i.e.* whorl, loop and arch.

Moreover, other machine learning techniques could also be investigated to strengthen the fact that naïve Bayes algorithm is suitable for discriminating sex based on diagonal ridge counts.

Acknowledgment

L.C. Lee thanks CRIM Universiti Kebangsaan Malaysia (Gant no. GUP-2019-085). The authors also like to extend thanks to the assistant laboratory staffs, especially Mr. Ab Aziz Ishak for their support in providing all the necessary tools needed in collecting the fingerprint data.

References

[1] M.H. Houck, Ed., Forensic Fingerprints, Amsterdam, Elsevier, 2016.

[2] B. Turvey and S. Crowder, Forensic Investigations, Academic Press: Elsevier, 2017.

[3] K.Q. Schulte, F.C. Hewitt, T.E. Manley, A.J. Reed, M. Baniasad, N.C. Albright, M.E. Powals, D.S. LeSassier, A.R. Smith, L. Zhang, L.W. Allen, B.C. Ludolph, K.L. K.L. Weber, A.E. Woerner, M.A. Freitas and M.W. Gardner, "Fractionation of DNA and protein from individual latent fingerprints for forensic analysis," Forens. Sci. Int.: Genetics, vol. 50, pp. 102405, 2021.

[4] C. Lennard, Forensic Sciences | Fingerprint Techniques, In Encyclopedia of Analytical Science, 3rd ed., pp. 38-47, 2019, Elsevier.

[5] N. Singla, M. Kaur and S. Sofar, "Automated latent fingerprint identification system: A review," Forens. Sci. Int., vol. 309, pp. 110187, 2020.

[6] A.I. Awad, Machine Learning Techniques for Fingerprint Identification: A Short Review, In Ell Hassanien A, et al. (Eds), AMLTA 2012, CCIS 322, pp. 524-531, 2021, Berlin Heidelberg, Springer.

[7] K. Noor, T. Jan, M. Basheri, A. Ali, R.A. Khalil, M.H. Zafar, M. Ashraf, M.I. Babar and S.W. Shah, "Performances enhancement of fingerprint recognition system using classifiers," IEEE Access, vol. 7, pp. 5760-5768, 2018.

[8] J.S. Yadav, A.K. Gupta and A. Saxena, "A review on gender identification using machine learning based on fingerprints," J. Inf. Opt. Sci., vol. 40, pp. 1121-1129, 2019.

[9] J.S. Yadav and A. Saxena, "A review on gender identification using machine learning technologies based on fingerprints," J. Biometrics & Bionstat., vol. 9, pp. 1000412, 2018.

[10] M.A. Acree, "Is there a gender diference in fingerprint ridge density?" Forens. Sc. Int., vol. 102, pp. 35-44, 1999.

[11] K. Maninder, K. Mankamal and Y. Jigmath, "Identifiction of sex using discriminant function analysis of fingerprint ridge density at three topological areas among North Indian population," Anthropol. Rev., vol. 83, pp. 349-361, 2020.

[12] F.I. Ali and A.A. Ahmed, "Sexual and topological variability in palmprint ridge density in a sample of Sudanese population," Forensic Sci. Int.: Reports, vol. 2, pp. 100151, 2020.

[13] M. Kaur, "Fingerprint Ridge Density of Convicted Male and Female Prisoners: A Pilot Study," Braz. J. Forensic Sci., Med., Law, and Bioeth., vol. 8, pp. 226-234, 2019.

[14] V.C. Nayak, P. Rastogi, T. Kanchan, K. Yoganarasimha, G.P. Kumar and R.G. Menezes, " Sex differences from fingerprint ridge density in Chinese and Malaysian population," Forens. Sci. Int., vol. 197, pp. 67-69, 2010.

[15] S.F. Abdullah, A.F.N.A. Rahman, and Z.A. Abas, "Classification of gender by using fingerprint ridge density in northern part of Malaysia," APRN J. Eng. Appl. Sci., vol. 10, pp. 10722-10726, 2015.

[16] A. Mishra, S. Dubey and A. Sahu, "Gender classification based on fingerprint database using association rule mining," Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Advances in Intelligent Systems and Computing, vol. 1245, pp. 121-133, 2020. https://doi.org/10.1007/978-981-15-7234-0_10

[17] P. Terhorst, N. Damer, A. Braun and A. Kuijper, "Deep and Multi-algorithmic gender classification of single fingerprint minutiae," 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, pp. 2113-2120, 2018. doi: 10.23919/ICIF.2018.8455803

[18] S. Suwarno and P.I. Santosa, "A short review of gender classification based on fingerprint using wavelet transform," Int. J. Adv. Comput. Sci. Appl., vol. 8, pp 562-564, 2017.

[19] O.N. Iloanusi and U.C. Ejiogu, "Gender classifiction from fused multi-fingerprint types," Infor. Secur. J.: A Glob. Perspect., vol. 29, pp. 209-219, 2020.

[20] S.F. Abdullah, A.F.N.A. Rahman, Z.A. Abas and W.H.M. Saad, "Development of a Fingerprint Gender Classification Algorithm Using Fingerprint Global Features," Int. J. Adv. Comput. Sci. Appl., vol. 7, pp. 275-279, 2016a.

[21] S.F. Abdullah, A.F.N.A. Rahman, Z.A. Abas and W.H.M. Saad, "Fingerprint gender classification using univariate decision tree (J48)," Int. J. Adv. Comput. Sci. Appl., vol. 7, pp. 217-221, 2016b.

[22] S.F. Abdullah, A.F.N.A. Rahman, Z.A. Abas and W. H. M. Saad. "Support Vector Machine, Multilayer Perceptron Neural Network, Bayes Net and k-Nearest Neighbour in Classifying Gender using Fingerprint Features," Int. J. Comput. Sci. Inf. Secur., vol. 14, pp.336-340, 2016c.

[23] S.F. Abdullah, A.F.N.A. Rahman, Z.A. Abas and W. H. M. Saad. "Multilayer Perceptron Neural Network in Classifying Gender using Fingerprint Global Level Features," Ind. J. Sci. Techn., vol. 9, pp. 1-6, 2016d.

[24] S. Alam, Dipti, M. Dua and A. Gupta, " A comparative study of gender classification using fingerprints," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 880-884, 2019.

[25] E.B. Ceyhan and S. Sagiroglu, "Gender inference within Turkish population by using only fingerprint feature vectors," 2014 IEEE Symposium on Computational Intelligence in Biometrics and Identify Management (CIBIM), 2014.

[26] L.C. Lee and M.Y. Adam, "A study of sex differences from fingerprint ridge density in the Malay population of the peninsular

Malaysia," National Forensic Science Symposium 2015, Pullman Hotel, 2015.

[27] L.C. Lee and N.I. Bohari and S.N. Sanih, "Gender discrimination based on ridge density in Malaysian Chinese population," Forensic Science Seminar 2018, Pullman Hotel, 2018.

[28] L.C. Lee, N.I. Bohari and S.N. Sanih, "Gender discrimination based on number of ridge in Malaysian Indians population," National Forensic Science Symposium (NFSS) 2019, Pullman Hotel, 2019.

[29] R Core Team, R, A language and environment for statistical computing, R Foundation for statistical computing, Vienna, 2017, http://www.R-project.org/.

[30] R. Wehrens and J. Kruisselbrink, Supervised and Unsupervised Self-Organising Maps, R package version 3.0.10, 2019, https://cran.r-project.org/web/packages/kohonen/kohonen.pdf

[31] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C-C. Chang and C-C. Lin, Misc Functions of the department of statistisc, probability theory group (Formerly: e1071), R package version 1.7-3, 2019, https://cran.r-project.org/web/packages/e1071/e1071.pdf

[32] T. Therneau, B. Atkinson an B. Ripley, Recursive partitioning and regression trees, R package version 4.1-15, 2019, https://cran.r-project.org/web/packages/rpart/rpart.pdf

[33] M.A. Olsen, E. Tabassi, A. Makarov and C. Busch, "Self-organizing maps for fingerprint image quality assessment," 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 10.1109/CVPRW.2013.28.

[34] A.N. Ouzounoglou, T.L. Economopoulos, P.A. Asvestas and G.K. Matsopoulos, "Fingerprint matchng with self-organizing maps," P.D. Bamidis and N. Pallikarakis (Eds). MEDICON 2010, IFMBE Proceedings 29 pp. 307-310, 2010.

[35] T. Kohonen, Self-organization maps, Heidelberg, Springer, 2nd ed. 1997.

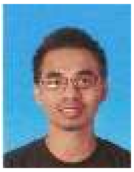
[36] L. Breiman, J. Friedman and R. Olshen, Classification and regression trees, California: Wadsworth Belement, 1984.

[37] D.J. Hand and K. Yu, "Idiot's Bayes-not so stupid at all?" Int. Stat. Rev., vol. 69, pp.385-398, 2001.

[38] A.Z. Mundorff, E.J. Bartelink and T.A. Murad, "Sexual dimorphism in finger ridge breath measurements: a tool for sex estimation from fingerprints," J. Forensic Sci., vol. 59, pp. 891-897, 2014.

[39] E. Gutierrez-Redomero, C. Alonso, E. Romero and V. Galera, "Variability of fingerprint ridge density in a sample of Spanish Caucasians and its application to sex determination," Forens. Sc. Int., vol. 180, pp. 17-22, 2008.

Loong Chuen Lee is a senior lecturer in the Forensic Science Program, Faculty of Health Sciences, Universiti Kebangsaan Malysia (UKM) and member of the Malaysian Institute of Chemistry (Institut Kimia Malaysia, IKM). IKM is a statutory professional organization incorporated under the Chemist Act 1975. Her current research is focused on developing analytical approaches for classification, discrimination and identification problems in the context of forensic sciences by using chemometric and chemical instrumental techniques.



Muhammad Yusuf Adam received her Bachelor of Sciences (Forensic Sciences) degree from Universiti Kebangsaan Malaysia, Malaysia in 2014.

Siti Norfaraan Abd Sanih received her Bachelor of Sciences (Forensic Sciences) degree from Universiti Kebangsaan Malaysia, Malaysia in 2019.

Nur Izzati Bohari received her Bachelor of Sciences (Forensic Sciences) degree from Universiti Kebangsaan Malaysia, Malaysia in 2019.