Journal of ELECTRICAL ENGINEERING, VOL 74(2023), NO1, 23–31

sciendo

PAPERS

A flexible approach for biometric menagerie on user classification of keystroke data

Mehmet Erdal Özbek[1]

Biometric systems aim to provide reliable authentication and verification of users. The behaviour of the users may alter the authentication performance when accessing these systems. Therefore, clustering users based on their actions is crucial. A biometric menagerie defines and labels user groups statistically according to their variability. However, determining groups is a fuzzy process and it may lead to inconsistencies. In this work, a novel and flexible approach is introduced based on the classification performance of the users data collected in a database without imposing any other restrictions. According to the performance measures obtained from the confusion matrix of the classification algorithms, users are ranked and then clustered. Additionally, the norm of a confusion matrix is offered augmenting the state-of-the-art performance metrics. The proposed scheme is evaluated using the behavioural biometrics modality on two benchmark keystroke databases. The performance results successfully illustrate the alternative way of grouping and identification of users sharing the same behaviour irrespective of the chosen classifiers or performance metrics.

K e y w o r d s: biometric menagerie, confusion matrix, keystroke dynamics

1 Introduction

Recognizing the identity of a person based on the physical or behavioural traits of that individual is commonly referred as biometrics. It is performed by an authentication process either for identification of the users or verifying users that claim to be that individual [1]. In accessing to the systems handling biometric data, either for identification or validation purposes, a binary decision of acceptance or rejection is compulsory. For performing any of the two objective, the system is usually designed to work at two separate stages: enrolment and verification. In the enrolment phase, the biometric information of the individuals is recorded and stored. In the verification phase, following the collection of the new input, the decision is made based on the pre-stored enrolment data to determine whether the recent input is from a genuine user or from an impostor.

In order to develop reliable biometric authentication, the system performance is generally evaluated using matching scores by comparing the query features extracted from the user data against the stored templates. Those measures reveal the degree of similarities due to intra-user and inter-user variations where the former originates from the variations of an individual, whereas the latter is due to different individuals. In a biometric system both variations occur although they are not evenly distributed across the users [2].

Based on those inherent differences in performance measures, a statistical framework has been developed to group users firstly in the context of speaker recognition [3]. The so-called Doddington zoo is a biometric menagerie [4] that defines and labels user groups according to some animal species reflecting their behaviour within the biometric systems. In this model, the user groups are represented by sheep, goats, lambs, and wolves. Sheep characterize the majority who have low false accept and low false reject error rates. The goats have high intra-user variations therefore they are difficult to recognize. Lambs correspond to users having high inter-user similarities and they are easily imitated. Wolves represent users who imitate the other users and thus increase the false acceptance rate [2,3].

Following the effectiveness in identification of users, a second menagerie has been offered by Yager and Dunstone [5], where new classes of animals have been defined concerning the relationship between the genuine and impostor scores. They introduced doves, chameleons, phantoms and worms, all characterized by the scores and the relationships between the animals. They have built combinations of low/high and genuine/impostor match scores leading to a group-centric approach. This is followed by other grouping schemes based on the clue that different images of the same subject might exhibit different matching rates, defining image-specific error rates [6]. The performance variations of the images in the biometric zoo are then reflected in categorizations as blue wolves, clear ice, blue goats, and black ice according to their level of recognizability.

The biometric menagerie has been considered to be fuzzy and inconsistent for iris recognition, so that the fuzzy-linguistic labels of menagerie in terms of first/last wolf-, sheep-, lamb-, goat- templates, all claimed to depend on the calibration of the recognition system [7].

[1] Department of Electrical and Electronics Engineering, Izmir Kâtip Çelebi University, Izmir, Türkiye, merdal.ozbek@ikcu.edu.tr

https://doi.org/10.2478/jee-2023-0003, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

However, formulating a user-specific score normalization scheme has been studied to analyse the performance variability [8]. That work has been extended to user-specific performance evaluation schemes by clustering users through a biometric menagerie index [9]. Similarly, personal entropy measure has been examined to quantify directly on genuine handwritten signatures for exploiting biometric menagerie [10]. Style signatures have been shown to handle the challenges posed by the goats/wolves/lambs [11]. A further categorization for the multi-biometric system has been developed with biometric selective fusion where a user might be separated into two categories of so-called well-behaved and weak [12].

Those works and further extensions mostly consider the legitimate and impostor scores. A recent study has employed an adaptive update mechanism improving the verification performance in an intra-class variation problem [13]. It has been shown that the use of classifications as a posteriori information enhances the performance. It is natural that some of the users are more difficult to recognize while some of them are easier. As biometrics is often data driven, it is straightforward to use classifiers to label users in the dataset. Particularly for the behavioural biometrics modalities, machine learning tools may be used to specify a user, based on the classification performance of users. The performance of the biometric system describes how well a user is recognized and differentiated from the other users [14]. However, the variations depend on not only the intrinsic differences between users [15] but also different samples obtained from the same user. The novel idea of this paper is based on the paradigm that evaluates the user data based on their consistency in achieving similar performance. Therefore, our aim is to cluster the users similar to a menagerie based on their ranking obtained from their classification performance. For that purpose, different classifier algorithms are used to evaluate the performance measures of the users while the existing studies rely on only the acceptance/rejection rates. In order to demonstrate the proposed approach, a behavioural biometrics modality, keystroke dynamics is preferred including many samples of intra- and inter-user variations.

2 Literature review

2.1 Menagerie models

Doddington menagerie clusters users according to their matching score in verification. It is designed by statistical procedures such as F-test, Kruskal-Wallis test, and Durbin test applied for variance analysis [3]. The groups, each named by an animal that users belong to, have been formed based on users behaviour. The selection of users belonging to either of the animal categories is based on finding the average match score for each user. The match score is evaluated with a standard hypothesis testing representing how likely the given data fits the model. The ratio of the expected probability that the two samples from the same user that is falsely stated as non-match gives the false non match rate (FNMR). Correspondingly, when the two samples from different users declared falsely as a match refers to the false match rate (FMR). In the context of biometric verification, FNMR and FMR are also known as false reject rate (FRR) and false accept rate (FAR), respectively. The system makes a binary decision based on a fixed threshold. If the score is higher than the threshold the hypothesis is accepted, otherwise rejected. As stated in [15], users can always be sorted in terms of average match scores and the bottom 2.5% might be labelled as goats. The similar approach is then considered for lambs and wolves based on the non-match scores. The remaining users are labelled as sheep [3,15].

Then, the idea of menagerie has been extended to categorize users according to some importance criteria. The Doddington biometric animals are based on only genuine or impostor match scores. The new animals introduced in [5] offers that a relationship between those scores can be found. A statistical information might be gathered based on some performance indicators displaying how well a group of users match against the users in the group and how well they match against the rest. This led to another interpretation that a group of users and their relations might be discovered. For this purpose, some performance measures can be chosen depending on the type of the considered biometric system. For example, counting the number of errors, ranking based on a performance criterion, score based results with maximum scores, minimum scores, or mean scores.

An important question arises whether the recognition performance of users is affected primarily due to the intrinsic differences of users or not. This has led to a framework for a hierarchy of menageries generalized with respect to the algorithms and the data sets [15]. The standard statistical hypothesis testing approach is used as the basis for describing and testing the existence of different levels of biometric zoo, such as zeroth-order, first-order and higher-orders of menagerie. A zeroth-order zoo is referred to that users may be labelled as animals in a single experiment. A first-order zoo exists when the user identity matters for other data drawn from the same scenario, and higher-order zoos display higher generalization.

Due to variations of the users while accessing to biometric systems, a ranking criterion has been offered to display the recognizability of the users based on their strength of performance [8]. A constrained F-norm ratio has shown to exhibit the best generalization ability related to the equal error rate (EER) where FRR and FAR values are equal. The ranking has been also used for grouping users demonstrated for image data according to their easiness of recognition [6]. A uniqueness measure score based on the Kullback-Leibler divergence has been described to quantify the identification capacity of faces by investigating the impact of feature extractors of deep neural network (DNN) algorithms [16]. Similar to the other biometric types, image performance variation is often related to the capturing devices or sensors, to

the environment conditions, or to the user. In either case, the recognition algorithms that bound FNMR and FMR values are more reliable and secure. Thus, finding other ways of restricting those error rates is the subject of further research.

2.2 Keystroke dynamics

Keystroke dynamics is recently popular but at the same time it is one of the oldest biometrics modality to identify users that use a typing device. The identification of a person based on how they use that kind of device is highly reasonable since the typing characteristics are distinctive enough to distinguish a user from another. This phenomenon has been already known from the early recordings of the telegraph era [17]. The important data captured from the dynamics of writing began within telegraphs, later transformed to typewriting and nowadays to the computers and smart devices with keyboards. The dynamics information is related to the changes in consecutive keystrokes, basically composed of key down and key up events which each individual performs unconsciously during typing [1,18]. The dynamics are mostly related to time domain information represented by the time differences between: press-press (PP), release-release (RR), press-release (PR), and release-press (RP) events as displayed in Fig. 1. The keystroke duration is known as hold time or dwell time, and the latency between successive keystrokes is referred as delay time. Since many decades, this behavioural biometrics modality has been included for authentication of users [17-23]. Recent works also include machine learning and deep learning techniques [24,25].

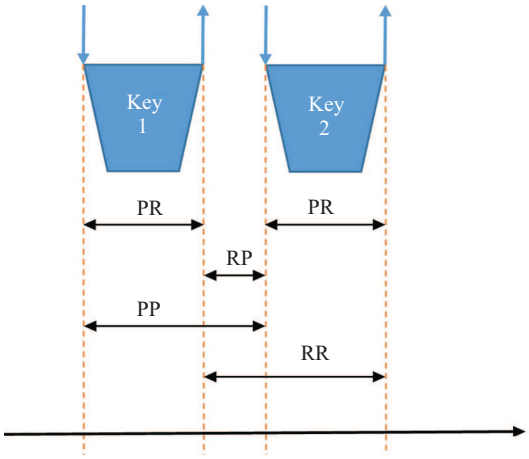

Fig. 1. Representation of keystroke dynamics time information data with press (P) and release (R) events of the keys

2.3 Performance measures

In biometric systems, evaluation is based on finding matching scores. The genuine and impostor match score distributions lead to FNMR and FMR values with a given threshold. The receiver operating curves (ROC) curves can be used to display the variations and the area under the ROC curve (AUC) is used to compare the performance of two different biometric systems. A single-valued common measure for summarizing the performance of a biometric system is the EER [2].

Similar evaluation for the classifier performance of a binary classification system is obtained based on distinguishing the actual class and the predicted class. Then four possible conditions are given as the true positive (TP) where the positive matches that are correctly classified, true negative (TN) where the negative matches that are correctly classified, false positive (FP) where the negative matches that are incorrectly classified as positive, and false negative (FN) where the positive matches that are incorrectly classified as negative. The most common performance measures based on these values are simply accuracy (A), precision (P), recall (R), and F-score metrics

$$A = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{TN}} + N_{\mathrm{FN}}}, \tag{1}$$

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \tag{2}$$

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}, \tag{3}$$

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}. \tag{4}$$

They are frequently used to evaluate the performance of the classification algorithms. Note that the F-score can be adjusted according to the value of β relating the precision and recall values. In general, $\beta = 1$ chosen to reflect the balanced importance of recall and precision and thus commonly abbreviated as F_1.

3 Methodology

3.1 Perspective

Doddington or other ensued menageries aim to cluster users belonging to a group named by an animal according to their performance measures mostly based on their matching score in verification. As the amount of data and the corresponding databases for biometric systems tend to increase, the use of machine learning algorithms are very likely to be helpful in determining and labelling users. Based on this idea, clustering users according to their classification performance from the data available in the databases constitutes the alternative way and the core of the proposed model.

The proposed model uses classification performance scores of the users and sorts them accordingly to obtain a ranking. The performance is deduced from the confusion matrix (also called as contingency table) computed for every user for each classification. The correct classification ratios are commonly reflected as accuracy scores. However, according to a defined menagerie finding the mismatched users based on how they are wrongly labelled as another user is the key for identifying the users and

their corresponding label. Those data lay mainly at the off-diagonal terms of the confusion matrix. Therefore, in order to reveal the differences of the users, we propose to emphasize the information collected from the off-diagonal terms of the confusion matrix. From a confusion matrix $C \in R^{m \times n}$ with elements $c_{i,j}$ where $i = 1, 2, \ldots, m$ and $j = 1, \ldots, n$, the modified confusion matrix is obtained

$$C_{\text{mod}} = \begin{cases} 0, & \text{if } i = j \\ c_{i,j} & \text{if } i \neq j \end{cases}. \tag{5}$$

The importance of those off-diagonal terms in classification performance can be easily visualised via an example confusion matrix in Fig. 2. It is seen that omitting the higher values at the diagonals reveal the misclassifications or confusions at the off-diagonals, thus helping user identification.

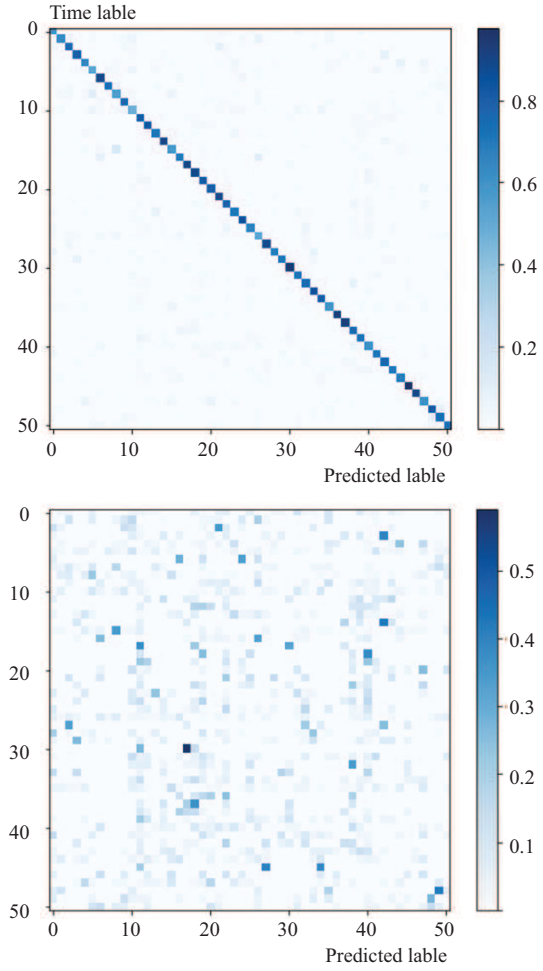

Fig. 2. An example confusion matrix and its modified version

Since the important information resides in the confusion matrix, in this work, we further propose to use the norm of the confusion matrix as augmenting the evaluation metrics. Thus, in order to demonstrate the dimension of confusion in identification, a metric based on the Frobenius norm of the confusion matrix is

$$||C||_{\text{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |c_{i,j}|^2}, \tag{6}$$

where F denotes the Frobenius norm. The value of this metric is directly proportional to the number of misclassifications and can be computed using the original confusion matrix as well as the modified confusion matrix C_{mod}. This value is named as confusion matrix norm (CMN).

Then for each user, the number of misclassified entries are retrieved from the confusion matrix. For each of the classifiers, the users from the highest performance metric to the lowest are sorted. In that list, both the highest and the lowest pre-defined percentile of the total number of the users are stored. After finding those users labelled with their identification numbers (IDs), the users who are common in each classifier are detected by fusing the information through the classifiers and the performance metrics. Therefore, based on the selected percentile of users, users can be clustered into three groups having high performance, low performance, and the remaining ones as middle performance. The scheme is flexible since any classification algorithm can be used, by easily determining the percentile of users, and choosing any of the metrics to compare with the other.

The pseudo-code for the user ID ranking fusion algorithm is provided below:

Algorithm 1. User ID ranking fusion algorithm

input: Keystroke data, percentile value
initialize: Classifiers, classifier parameters
repeat
  for each user
    Compute confusion matrix and obtain TP, TN, FP, FN values
  endfor
  Compute and sort A, P, R, F1, and CMN scores
  for each metric
    Store the user IDs for H, M, L groups
  endfor
until Total number of classifiers
Fusion of classifier outputs to determine the user IDs common for each classifier
display: User IDs in any two selected metrics

3.2 Data

In this work, two keystroke benchmarking databases are considered to demonstrate the performance of the proposed menagerie model. A brief explanation of the databases is given as following.

GREYC: It consists of 7555 captures from 133 individuals. From an AZERTY keyboard users typed *greyc laboratory*. For each of the typing record, the data contain the code of the key, the type of the event, and the time of the event. The information of time differences between PR, RR, PR, RP events, with an additional vector

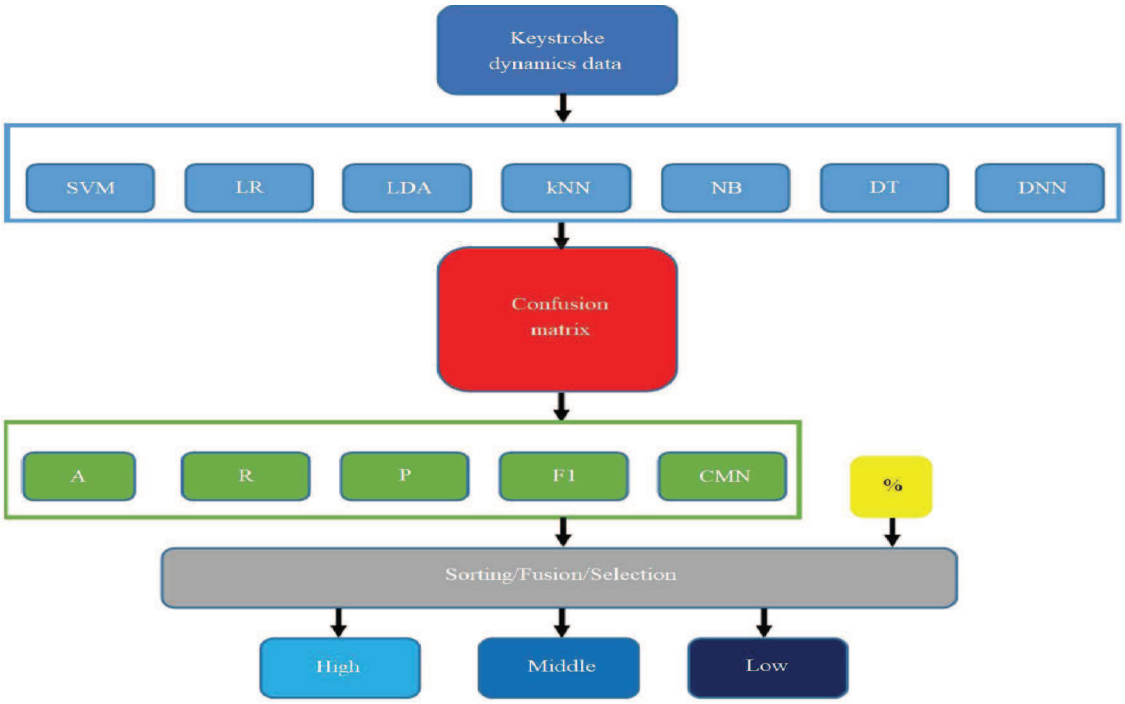

Fig. 3. Summary of the proposed model

of the concatenation of the previous ones, and the total typing time of the passphrase have been stored in the database [26].

CMU: It consists of timing information from 51 individuals, each typing 400 times the password *.tie5Roanl* with a QWERTY keyboard. For each password, the *Enter* key has been considered to be a part of the password and the PP, PR, and RP time differences have been extracted for all of the keys and then collected into a single vector [27].

Since the collected vectors are directly used in classifications for both datasets, any pre-processing or feature selection step is not performed.

3.3 Classification methods

The classification of users data can be performed by various machine learning algorithms as well as deep learning architectures. For this study six different classifiers and one DNN model are utilized to demonstrate the proposed scheme. The selected classifiers are support vector machine (SVM), logistic regression (LR), linear discriminant analysis (LDA), K-nearest neighbourhood (KNN), naive Bayes (NB), and decision tree (DT). In case of DNN, a simple architecture is selected for the ease of comparison. A brief information for each of them is given as following.

SVM is one of the commonly used supervised learning model for binary classification problems. A hyperplane is found between the classes that separates the classes with a maximum margin. The data that are close to the boundary hyperplane are the support vectors that allow to learn the discriminant function. High dimensional features can be easily classified with higher accuracy rates.

The LR classifier provides the probability measures to determine the binary output. It describes a conditional distribution where posterior probability of a class can be written as a logistic sigmoid function. The parameters of the model are found by the maximum likelihood algorithm.

LDA is a supervised method for dimension reduction. It determines the direction of the data when projected onto a (weight) vector in order to separate the examples of the two classes as well as possible.

KNN is a simple and efficient algorithm that selects the class based on the majority vote of K closest points. It is based on the idea that the more similar or closer the instances, the more likely that they belong to the same class. The closeness is measured by a distance or similarity measure between the data samples.

In NB classifiers, the attributes are assumed to be independent of each other. They are conditioned on class labels and mostly assumed to have Gaussian distributed. A DT is a hierarchical model for supervised learning where a sequence of binary selections forms a tree structure that traversing from features to classification outcome is processed. On the other hand, DNN models demonstrated higher performance in machine learning problems with various number of architectures. From a simple multilayer perceptron (MLP) model to more complex convolutional neural networks (CNN) and recurrent neural networks (RNN), many models are available. They depend on the use of hidden layers where they differ in extracting fea-

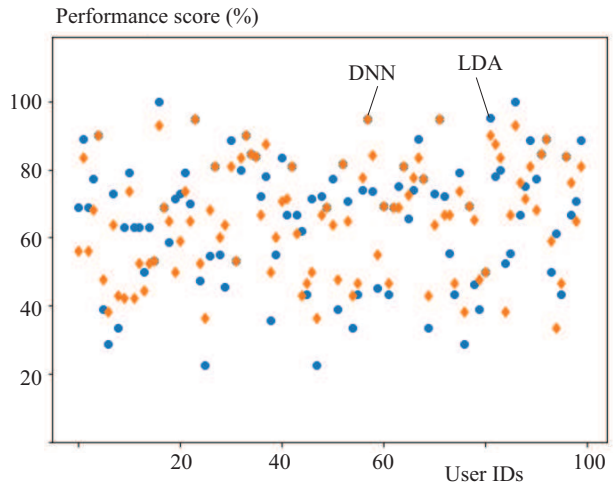

Fig. 4. The performance comparison of the two classifiers for each user

Table 1. Classifier performance results for GREYC data

Classifier	A	P	R	F_1	CMN	CMN_{off}
SVM	0.66	0.68	0.66	0.66	0.93	0.98
LR	0.59	0.59	0.60	0.59	0.90	0.90
LDA	0.69	0.71	0.69	0.69	0.93	0.99
KNN	0.60	0.62	0.60	0.59	0.93	0.98
GNB	0.53	0.52	0.52	0.51	0.94	0.98
DT	0.48	0.48	0.48	0.47	0.94	0.98
DNN	0.67	0.68	0.67	0.66	0.93	0.99

Table 2. Classifier performance results for CMU data

Classifier	A	P	R	F_1	CMN	CMN_{off}
SVM	0.77	0.79	0.77	0.77	0.89	0.99
LR	0.39	0.56	0.39	0.34	0.92	0.96
LDA	0.73	0.74	0.73	0.73	0.89	0.99
KNN	0.71	0.73	0.71	0.71	0.90	0.97
GNB	0.59	0.61	0.59	0.58	0.91	0.98
DT	0.63	0.63	0.63	0.63	0.91	0.99
DNN	0.87	0.87	0.87	0.87	0.88	0.99

tures using convolutions or using sequential data with varying number of layers and layer-connections.

3.4 Proposed model

The implementation of the proposed model can be summarized as following. First, the behavioural biometrics data, that is, the keystroke data from the databases summarized in Section 3.2 are retrieved. In this paper, we have used 100 users from GREYC data based on the work that offers to discard users having less number of acquisitions [28]. Then the classification performance of each user is obtained from different classifiers provided in Section 3.3 in terms of performance metrics based on the confusion matrix. The fusion of information based on the sorting of classification performance values with a pre-defined percentage level leads to grouping of users as high, middle, and low as in Algorithm 1. The middle group is defined as the users having performance percentages other than separated as high or low.

A graphical representation of the proposed model is presented in Fig. 3. Firstly, using the keystroke dynamics data, different classifiers are utilized to obtain a confusion matrix where misclassification results are collected. For each confusion matrix, performance metrics provide the selected scores. Then, the main algorithm sorts, combines, and selects the higher and lower range of these scores based on the pre-defined percentage level illustrated by %. Finally, the algorithm ends with selection of performance levels.

4 Performance results

In this section, we present performance results of the proposed scheme using the two aforementioned databases, GREYC and CMU. For all of the experiments performed in this work, the data are divided into training and testing groups randomly with 70% and 30%, respectively. The classifications are performed and the confusion matrix is obtained for each classifier utilized with their default parameter values. A linear classifier is selected for the SVM classifier. The number of neighbours, K, is chosen as 5 for KNN. No grid search or optimization of the hyperparameters is performed. For the DNN, four layers including input layer with 128, 256, 256, and an output layer based on the number of users in the database is used. The rectified linear unit (ReLU) activation function is used at the input and at the middle layers where a sigmoid function is chosen at the output layer. Binary crossentropy function is selected as the loss function with the Adam optimizer. The state-of-the-art performance metrics, *ie* accuracy, precision, recall and F1 scores are computed. Moreover, the proposed metrics computed from the confusion matrix, *ie* CMN and its equivalent without the diagonal terms in the confusion matrix (labelled as CMNoff) are included. They are normalized to unity and subtracted from unity to display similar range with the other performance metric scores.

The most important parameter of the proposed method is the percentage value that is used to make groupings. While it is an adjustable parameter for our algorithm, we kept it as 2.5% as in [15].

4.1 Results for GREYC data

Table 1 displays the average classification performance of the selected classifiers with respect to the computed metrics. The LDA classifier has found to have the highest values in terms of accuracy, precision, recall and F1 scores. On the other hand, the error measure CMN displays the worst performance for LDA that opposes the

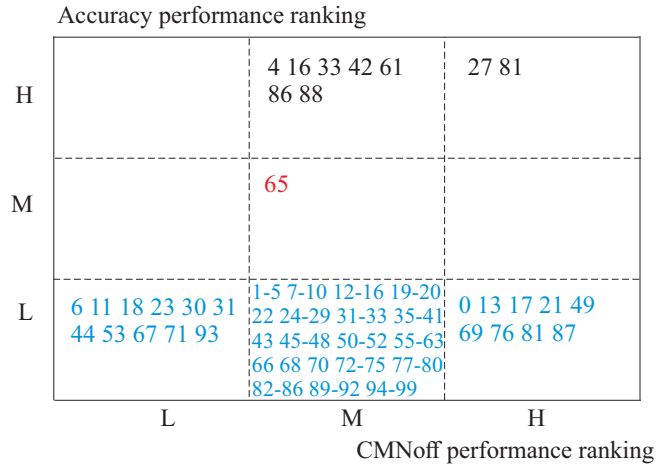

Fig. 5. User IDs grouped for accuracy versus CMNoff performance

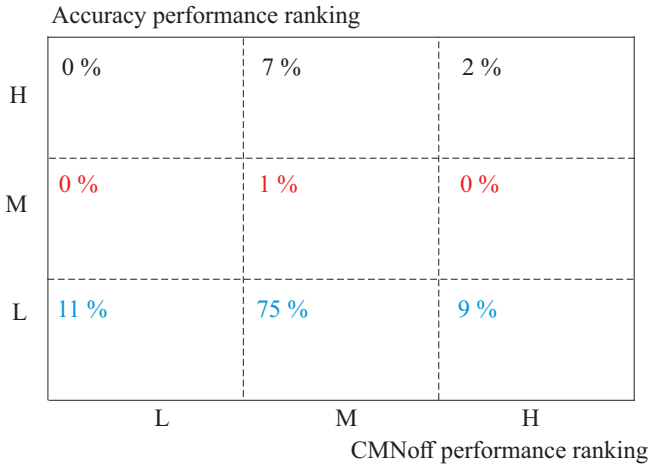

Fig. 6. Distribution of users grouped for accuracy vs confusion matrix norm performance

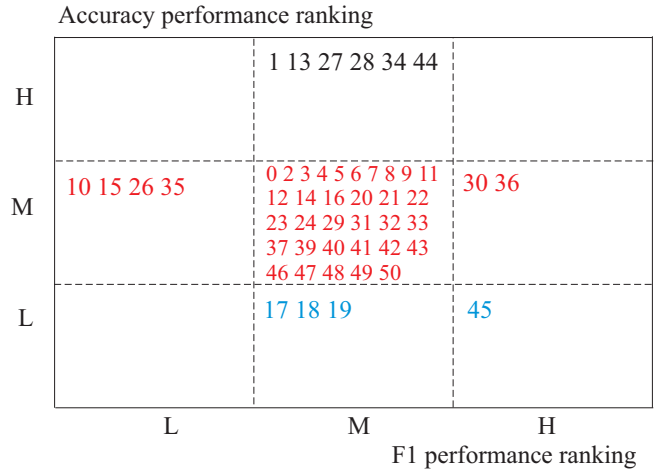

Fig. 7. User IDs grouped for accuracy versus F1 performance

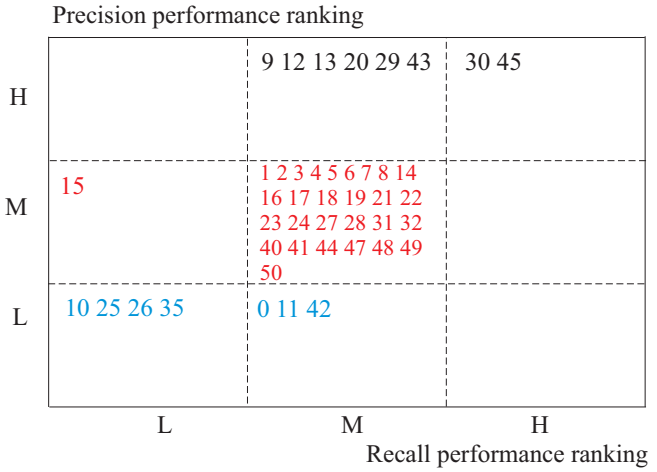

Fig. 8. User IDs grouped for precision versus recall performance

results with those scores. However, the proposed CMNoff metric fixes this wrong interpretation and shows similar performance as the other metrics reflecting appropriateness of using it.

As an example, the accuracy performance of the two classifiers for each user is displayed in Fig. 4. Note that most of the users have similar or close performance results. However, some of the users have poor performance results although the classifier performs well on the average as given in Tab. 1. Inherently, some of the users have superior results than the average. The idea of the proposed work reveals itself here. When the users are ranked according to their performances, they will have different rankings due to their performances in various classifiers. The similar rankings when compared within all of the classifiers will inform us how the user data is consistent and therefore the user can be identified based on self-performance.

The proposed fusion algorithm then groups users in low (L), middle (M), and high (H) categories according to their rankings. The flexible properties of the proposed algorithm give an opportunity to select any of the computed metrics to compare and determine the user IDs that belong to the nine possible categories.

According to the accuracy versus CMNoff performance results shown in Fig. 5, most of the users have lower accuracy values. However, some of the users can be identified when compared their accuracy performance to CMNoff. As shown in Fig. 5, the users with IDs 27 and 81 present higher performance values than the other users both in accuracy and CMNoff scores across all classifiers. In this case, users data is consistent, thus the users can be easily identified. The user group with low accuracy but displaying high performance for confusion matrix norm should be further investigated. Therefore, the proposed fusion mechanism shatters the users so that an unusual, particular, or peculiar user might be identified. Moreover, when the number of users is high, displaying all the user IDs might not be adequate. Besides, a more general view of the users might be more informative. Therefore, a simple representation of the distribution of users according to the performance levels can be given as displayed in Fig. 6 for the same data considered and displayed in Fig. 5.

4.2 Results for CMU data

Similar results are obtained from the classification, sorting, and then fusion processes for the CMU data. Tab. 2 displays the classification performance of the classifiers with respect to the computed metrics. This time the highest classification performance is obtained from the DNN classifier. As explained in Section 4.1, the CM-Noff metric adequately performed with the lowest error value correspondingly with the other metrics.

In order to demonstrate another performance measure comparison, the users grouped according to the accuracy versus F1 scores is presented in Fig. 7. Note that most of the users are grouped in the middle range. The users having high accuracy performances show middle F1 scores so that their scores may be debatable.

Another performance comparison may be visualized from the comparison of precision and recall performance rankings as presented in Fig. 8. The users having low performance in both measures can be identified directly. This is also valid for all the other user groups. For example, user 30 and user 45 are identified as the only users having high precision and high recall performances.

4.3 A summary and evaluation of the results

The proposed scheme, simply, is an efficient way of making a menagerie. As opposed to the existing structures, user groups are established based on their performance in classifiers and grouped as high, middle, and low. The displayed user IDs demonstrate that users can be identified for different metrics after fusing their performance rankings in various classifiers. The performance is based on the confusion matrix information where the users at the off-diagonal terms of the confusion matrix are the candidates that do not match with the users.

The similar results for the CMU data act as a proof that the method is not a consequence of the selected GREYC data and thus not limited to a single database but it can be applied to any database. As the number of users in a database vary, representing their IDs may not be adequate. Then a percentage information given as in Fig. 6 will be helpful to group users. The selection of the percentage of the users will directly affect the number of users in groups. The selected number of users can be adjusted according to the number of users in the database, or a fixed value can be chosen according to a selected precision, not necessarily to be 2.5%.

The work presented here do not impose any restrictions on the data or the features extracted from the data. They can be pre-processed according to the requirements and then are fed to the classifiers. This is also valid for the chosen classifiers. The classifiers selected for this study are representative and they are given to handle the options of selecting among many classifier schemes. Nevertheless, one can use less or more number of classifiers, more fine-tuned versions of classifiers, or tailored classifiers in order to make elaborations of the results. This may help to identify the users accordingly while the proposed scheme is able to support them all.

For the fusion algorithm, no preference is sought for the fusion of performance metrics or classifiers. Thus an equal weighting for the metrics/classifiers in fusion is considered. Any preference or weighting scheme can be imposed for either the performance metrics, the classifiers or both.

5 Conclusion

Biometric menageries aim to categorize users according to their recognizability based on the inherent differences of users. The behavioural traits like keystroke dynamics give clues to identify a user based on how they are similar or dissimilar compared to the others without any control over their acts. The similarity of users is generally computed by matching scores to decide if the user is genuine or an impostor. However, a decision threshold is necessary to evaluate how good the user can be identified. For different conditions, menageries then try to group users according to their matching and non-matching scores and their relations.

The ever-increasing number of users accessing to the biometric systems and the amount of data collected for this purpose require machine learning algorithms to effectively handle verification processes. Inspired by the previous menagerie models, in this work, we presented a novel efficient ranking/fusion model for behavioural biometrics data. The keystroke dynamics data stored in the databases are considered to build a menagerie based on the classification performance of various classifiers and metrics. The core of the model depends on the confusion matrix where a misidentified user can be easily detected by eliminating the diagonal terms. Then, the state-of-the-art performance measures are used for ranking of the users. Based on the intuitive information that the off-diagonal terms of the confusion matrix may lead to identify users, we further proposed a confusion matrix norm as a performance metric and demonstrated its effectiveness by showing that it performs similarly as the other metrics.

Performance results listed for user IDs in tables are representative in order to demonstrate the concept of grouping users according to their classification performance. Thus, our menagerie is composed of nine groups: three different groups as high, middle, and low for any two comparable metrics. The proposed scheme is flexible in terms of selecting any biometrics data, metrics, classifiers, or percentage of users, although the user groups are not labelled with animal species.

As a future work, the scheme can be extended with new clustering categories. The clustering can be also specialized in order to focus on a single animal with two or more level of strategies.

Building biometric menageries is still an open issue and it surely includes fuzziness. The major difference of this work compared to the previous studies is that it does

not depend on the matching scores but on the classification performances. While the classification performances and evaluation metrics may vary, we believe that this work reveals an alternative way of grouping users based on the fusion of classifications and metrics. Thus it may further lead to more research efforts to shed light on those gray-shaded areas.

References

[1] A. A. Ross and A. K. Jain, "Biometrics, Overview", *Li SZ, Jain AK (Eds.), Encyclopedia of Biometrics*, Springer, pp. 289-294, 2015.

[2] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*, New York, USA, Springer, 2011.

[3] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", *Tech. Rep*, National INST of Standards and Technology Gaithersburg MD, 1998.

[4] K. O'Connor and S. Elliott, "Biometric zoo menagerie", *Li SZ, Jain AK (Eds.). Encyclopedia of Biometrics*, Springer, pp. 1-4 https://doi.org/10.1007/978-3-642-27733-7_9146-2, 2014.

[5] N. Yager and T. Dunstone, "The biometric menagerie", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 220-230 https://doi.org/10.1109/TPAMI.2008.291, 2010.

[6] E. Tabassi, "Image specific error rate: a biometric performance metric", *International Conference on Pattern Recognition*, Istanbul, Turkey, pp. 1124-1127, 2010.

[7] N. Popescu-Bodorin, V. E. Balas, and I. M. Motoc, "The biometric menagerie - A fuzzy and inconsistent concept", *Balas V, Fodor J, Varkonyi-Koczy A, Dombi J, Jain L. (Eds.). Soft Computing Applications, Advances in Intelligent Systems and Computing*, Springer, Berlin, Heidelberg, 195, pp. 27-43 https://doi.org/10.1007/978-3-642-33941-7_6, 2013.

[8] N. Poh, S. Bengio, and A. Ross, "Revisiting Doddingtons zoo: A systematic method to assess user-dependent variabilities", *Proceedings of Second International Workshop on Multimodal User Authentication*, Toulouse, France, pp. 1-7, 2006.

[9] N. Poh and J. Kittler, "A biometric menagerie index for characterising template/model-specific variation", *Tistarelli M, Nixon MS. (Eds.): International Conference on Biometrics, Advances in Biometrics*, LNCS 5558, pp. 816-827 https://doi.org/10.1007/978-3-642-01793-3_83, 2009.

[10] N. Houmani and S. Garcia-Salicetti, "On hunting animals of the biometric menagerie for online signature", *PLoS ONE*, vol. 11, no. 4, pp. e0151691 https://doi.org/10.1371/journal.pone.0151691, 2016.

[11] K. Sundararajan, T. J. Neal, and D. L. Woodard, "Style signatures to combat biometric menagerie in stylometry", *International Conference on Biometrics*, Gold Coast, QLD, Australia, pp. 263-269, 2018.

[12] A. Ross, A. Rattani, and M. Tistarelli, "Exploiting the Doddington zoo effect in biometric fusion", *Proceedings of 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems*, Washington DC, USA, pp. 1-7, 2009.

[13] A. Mhenni, E. Cherrier, C. Rosenberger, N. Essoukri, and B. Amara, "Analysis of Doddington zoo classification for user dependent template update: Application to keystroke dynamics recognition", *Future Generation Computer Systems*, 97, pp. 210-218 https://doi.org/10.1016/j.future.2019.02.039, 2019.

[14] D. Migdal, I. Magotti, and C. Rosenberger, "Classifying biometric systems users among the Doddington zoo: Application to keystroke dynamics", *Proceedings of the 18th International Conference on Security and Cryptography*, pp. 747-753, 2021.

[15] M. N. Teli, J. R. Beveridge, P. J. Phillips, G. H. Givens, D. S. Bolme, and B. A. Draper, "Biometric zoos: Theory and experimental evidence", *International Joint Conference on Biometrics*, Washington, DC, USA, pp. 1-8, 2011.

[16] M. Balazia, S. L. Happy, F. Bremond, and A. Dantcheva, "How unique is a face: An investigative study", *25th International Conference on Pattern Recognition*, Milan, Italy, pp. 7066-7071, 2020.

[17] S. Douhou and J. R. Magnus, "The reliability of user authentication through keystroke dynamics", *Statistica Neerlandica*, vol. 63, no. 4, pp. 432-449 https://doi.org/10.1111/j.1467 -9574.2009.00434.x, 2009.

[18] S. Banerjee, Z. Syed, N. Bartlow, and B. Cukic, "Keystroke recognition", *Li SZ, Jain AK (Eds.). Encyclopedia of Biometrics*, Springer, pp. 1067-1073 https://doi.org/10.1007/9781-4899-7488-4_205, 2015.

[19] R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro, "Authentication by keystroke timing: Some preliminary results", *Technical Report*, Rand Corporation, Santa Monica, CA, USA, 1980.

[20] R. Joyce and G. Gupta, "Identity authentication based on keystroke latencies", *Communications of the ACM*, vol. 33, no. 2, pp. 168-176 https://doi.org/10.1145/75577.75582, 1990.

[21] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication", *Future Generation Computer Systems*, vol. 16, no. 4, pp. 351-359 https://doi.org/10.1016/S0167-739X(99)00059-X, 2000.

[22] M. Karnan, M. Akila, and N. Krishnaraj, "Biometric personal authentication using keystroke dynamics: A review", *Applied Soft Computing*, vol. 11, no. 2, pp. 1565-1573 https://doi.org/10.1016/j.asoc.2010.08.003, 2011.

[23] R. Giot, M. El-Abed, and C. Rosenberger, "Keystroke dynamics authentication", *Yang J. (Ed.). Biometrics. InTechOpen*, pp. 157-182 https://doi.org/10.5772/17064, 2011.

[24] H. C. Chang, J. Li, C. S. Wu, and M. Stamp, "Machine learning and deep learning for fixed-text keystroke dynamics", *Stamp, M., Aaron Visaggio, C., Mercaldo, F., Di Troia, F. (eds) Artificial Intelligence for Cybersecurity. Advances in Information Security*, 54, Springer, Cham, pp. 309-329 https://doi.org/10.1007/978-3-030-97087-1_13, 2022.

[25] A. Acien, A. Morales, J. V. Monaco, R. Vera-Rodriguez and J. Fierrez, "TypeNet: Deep learning keystroke biometrics", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 57-70 https://doi.org/10.1109/TBIOM.2021.3112540, 2022.

[26] R. Giot, M. El-Abed, and C. Rosenberger, "GREYC Keystroke: A benchmark for keystroke dynamics biometric systems", *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, USA, pp. 1-6, 2009.

[27] K. S. Killourhy and R. A. Maxion, "Comparing anomaly detectors for keystroke dynamics", *Proceedings of the 39th Annual International Conference on Dependable Systems and Networks*, Estoril, Lisbon, Portugal, pp. 125-134, 2009.

[28] M. E. Özbek, "Classification performance improvement of keystroke data", *Innovations in Intelligent Systems and Applications Conference*, vol. no. ASYU, pp. 1-4 https://doi.org/10.1109/ASYU48272.2019.8946366, 2019.

Received 24 January 2023