Biometric Confusion Matrix and Inter ZooPlot: Two Novel Visualizations for Biometric Verification Evaluation

Boyu Zhu, Romain Giot
LaBRI, University of Bordeaux
boyu.zhu@u-bordeaux.fr, romain.giot@u-bordeaux.fr

Abstract

Biometric Verification Systems (BVS) often suffer from misclassification errors, which are frequently concentrated around a small subset of users for whom the system performs poorly. The Biometric Menagerie was introduced to categorize such users based on their biometric behavior. However, its main representation, the ZooPlot, fails to accurately distinguish all categories defined in the original Menagerie, particularly "Lambs" (easily impersonated) and "Wolves" (frequently impersonate others).

This paper proposes two new visualizations for evaluating BVS: Inter ZooPlot and the Biometric Confusion Matrix (BCM). A user study was conducted to assess the effectiveness of different visualizations. Inter ZooPlot and BCM achieved average accuracies of 90.0% and 89.4%, respectively, in distinguishing the four original categories defined in the Biometric Menagerie, outperforming the baseline ZooPlot at 73.9%. Furthermore, we show that BCM can reveal sources of user-specific errors and highlight system imbalances, making it a promising post-hoc explainability method for biometric system analysis. All additional materials are available at: https://github.com/Boyu1998/BCM.

1. Introduction

Biometric Verification Systems (BVS) are widely used in various applications, such as biometric passports, forensic investigations, and individual identification. With the widespread adoption of smartphones, people increasingly rely on these systems [27, 19]. However, they are prone to errors, making it essential to use evaluation methods that can help assess their performance [21].

Various evaluation metrics and visualizations have been developed to evaluate BVS and can be categorized into two levels: global and local [5]. Global-level evaluation methods evaluate the overall performance of the system, providing an abstract performance summary. Standard evaluation metrics, including Equal Error Rate (EER), False Match Rate (FMR), and False Non-Match Rate (FNMR), effectively capture the global performance of a BVS [10]. Conversely, local-level evaluation methods focus on evaluating the performance of users within the system. Since users exhibit varying levels of performance, each user may have distinct FMR, FNMR, and EER values. Researchers are particularly interested in identifying poorly performing user groups, as they may be responsible for a disproportionate number of verification errors [5].

Doddington *et al.* categorized users into four groups of animals, known as the Biometric Menagerie [4], based on their average genuine and impostor scores. Sheep match well with themselves and poorly with others. Goats are users who are difficult to match against themselves. Lambs are easily impersonated. Wolves are particularly successful at impersonating others. ZooPlot, a scatter plot of genuine and impostor scores, was proposed to visually distinguish user groups based on the statistical distribution of their scores [28]. However, ZooPlot doesn't specify whether an impostor score originates from an impersonator or the one being impersonated, which makes it unable to distinguish between Lambs and Wolves. This distinction is particularly important in biometric applications where the enrollment and verification processes differ significantly. For example, in real-world evaluations, enrollment may be conducted under controlled conditions using high-quality sensors, while verification often takes place in less controlled environments. Users who are frequently impersonated may indicate poor-quality enrollment templates, whereas users who frequently succeed in impersonating others may point to vulnerabilities in the verification stage [5].

Evaluation methods also hold potential for applications in the field of Explainable Artificial Intelligence (XAI). BVS mainly relies on machine learning algorithms, which are typically black-box models [9]. XAI aims to develop methods that can explain the outputs of these black-box models, providing human-understandable explanations for machine learning predictions and decisions. Currently, research on XAI in biometrics primarily focuses on the ex-

plainability of models in specific modalities, such as face recognition [1] and ocular recognition [13]. No prior research has explored providing model explainability from a post-hoc perspective in this context.

Our research introduces two visualizations for evaluating BVS. Inter ZooPlot was developed by extending ZooPlot to incorporate additional impostor score information, enabling a more detailed characterization of user behavior. Biometric Confusion Matrix (BCM) is inspired by the confusion matrix and directly visualizes verification error rates between all pairs of users. In our user study, Inter ZooPlot and BCM achieved a classification accuracy of 90.0% and 89.4% in distinguishing the four original categories defined in the Biometric Menagerie respectively, compared to 73.9% obtained with the baseline ZooPlot. Additionally, qualitative analysis across different datasets revealed that BCM exhibits potential as a component of post-hoc XAI methods. BCM can reveal latent clustering effects and detect anomalous users, offering deeper insights into system behavior beyond conventional metrics.

Section 2 discusses existing methods for evaluating BVS. Section 3 introduces Inter ZooPlot and BCM. Section 4 presents our experimental protocol and analyzes the experimental results. Section 5 concludes the work and proposes some points for future studies.

2. Related Works

This section first discusses how BVS are evaluated with commonly used evaluation metrics, then reviews existing metrics used to evaluate potential biases within BVS.

2.1. Verification Process

The verification process in BVS consists of two main phases: enrollment and verification. During enrollment, a user provides biometric data processed and stored as a biometric template to represent the user in the Gallery. In the verification phase, the user presents a newly collected sample of biometric data, referred to as a Query, which the system compares with the stored template of the claimed user. In an evaluation scenario, the set of all queries submitted for verification is called the Probe set [5].

The system generates a score representing the degree of match between the query and the template. A decision threshold is then applied to convert the score into a decision. To evaluate the system's ability to distinguish between genuine and impostor matches, it is common to match each sample in Probe against all users in the Gallery. This process generates a score dataset and a corresponding decision dataset across the entire population. These two datasets serve as the foundation for evaluating a BVS.

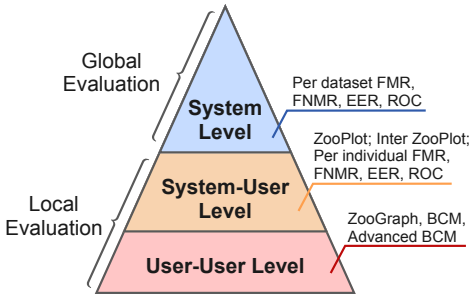


Figure 1. Biometric Evaluation Pyramid, which consists of three hierarchical levels. Deeper levels encode more detailed information than upper ones.

2.2. Evaluation Metrics and Visualizations

Evaluation metrics and visualizations for BVS can be categorized into two levels: Global and Local. Building upon this distinction, we further divide the Local level into two sub-levels: System-User Level and User-User Level, depending on whether the evaluation method reveals relationships between users. These three levels of evaluation can be summarized using a biometric evaluation pyramid, as illustrated in Figure 1.

System Level Evaluation provides an abstract representation of the overall system performance. Several standard evaluation metrics exist for BVS [10]. We typically evaluate the global performance based on the proportion of incorrectly accepted or rejected users. BVS errors have two categories: Type I error, where the system wrongly rejects a genuine user (False Negative), and Type II error, where the system mistakenly accepts an impostor (False Positive). Using these two error types, we derive two metrics to evaluate the overall performance of a BVS [23]. (a) False Match Rate (FMR), represents the proportion of impostor attempts that are incorrectly accepted. (b) False Non-Match Rate (FNMR), represents the proportion of genuine users who are wrongly rejected.

Another widely used evaluation visualization is the Receiver Operating Characteristic (ROC) curve, which plots FMR against Genuine Match Rate, i.e., the proportion of genuine attempts correctly accepted by the system, at different threshold values [18]. An evaluation metric derived from ROC is the Equal Error Rate (EER), which is the threshold of decision at which FMR equals FNMR. A lower EER indicates a more reliable BVS. The Area Under the Curve (AUC) of the ROC reflects the overall discriminative ability of the system: a higher AUC indicates better performance across all thresholds.

System-User Level Evaluation focuses on the performance of each user within the system. Users have different levels of performance within the same biometric system. To characterize such differences, Doddington *et al.* intro-

duced the Biometric Menagerie [4], categorizing users into four classes of animals based on their performance. Yager and Dunstone extended this taxonomy by introducing four new types: worms, doves, phantoms, and chameleons [28]. These new animals are defined not only by users' genuine or impostor scores, but by the relationship between the two. Doves match well with themselves and poorly with others, representing ideal users. Worms perform poorly in both genuine and impostor comparisons and are likely to cause the most system errors. Phantoms generate low scores regardless of who they are matched against. Chameleons appear similar to everyone, leading to high match scores in both genuine and impostor cases.

ZooPlot [28] is the evaluation visualization of this level, which maps users in a two-dimensional space according to their average genuine and impostor match scores. While ZooPlot effectively distinguishes the new categories, its reliance on aggregated impostor scores limits its ability to accurately identify the original four types defined by Doddington *et al.* [4].

User-User Level Evaluation focuses on the performance of each user within the BVS, as well as their relationships with other users in the system. Evaluation of this level has potential to support post-hoc explainability by revealing user-to-user relationships that may indicate systemic vulnerabilities. ZooGraph [6] is a graph-based evaluation visualization of this level, designed to reveal the relationships between users in BVS. Unlike ZooPlot, which primarily classifies users, ZooGraph displays the connections between users, providing deeper insights into abnormal matching behaviors within the system.

3. Proposed Methods

This section first introduces the distinction between different types of scores in Section 3.1. Then, Sections 3.2 and 3.3 respectively present two visualizations: Inter ZooPlot and the Biometric Confusion Matrix (BCM).

3.1. Types of Scores

The original Biometric Menagerie categories are determined based on the distribution of users' genuine and impostor scores. It is necessary to first define these two types of scores. Since impostor scores refer to the similarity scores involving user/impostor comparisons, we refer to them as Inter Scores in the following. Correspondingly, genuine scores will be referred to as Intra Scores.

To distinguish between Lambs and Wolves, it is necessary to define the different types of impostor scores. Let $S\left(G_i, P_j^k\right)$ denote the similarity score obtained from the k-th verification attempt in which user j from the Probe set is matched against user i from the Gallery, where $i \neq j$ ensures that the probe and gallery users are distinct. Let $\mathcal{U}$ be the set of all users, and let $N = |\mathcal{U}|$ be the total number of

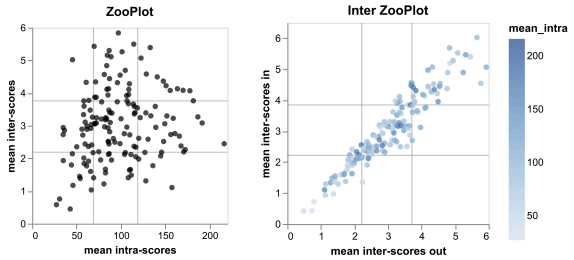


Figure 2. Comparison of ZooPlot and Inter ZooPlot. By incorporating additional information and transforming both axes, Inter ZooPlot provides a distinct analytical view of BVS.

users in the dataset. For each pair $(i, j) \in \mathcal{U} \times \mathcal{U}$, we let $M_{i,j}$ be the number of impostor verification attempts from user i to user j. We define the following impostor score sets for each user $i \in \mathcal{U}$:

$$\text{InterScoreInSet}(i) = \bigcup_{\substack{j \in \mathcal{U} \\ j \neq i}} \{S(G_i, P_j^k) \mid 1 \leq k \leq M_{i,j}\} \quad (1)$$

$$\text{InterScoreOutSet}(i) = \bigcup_{\substack{j \in \mathcal{U} \\ j \neq i}} \{S(G_j, P_i^k) \mid 1 \leq k \leq M_{j,i}\} \quad (2)$$

We now define the Intra Score using a set-based formulation. Let $\text{IntraScoreSet}(i)$ denote the set of genuine similarity scores obtained from successful verification attempts where user i is both the probe and the template:

$$\text{IntraScoreSet}(i) = \{S(G_i, P_i^k) \mid 1 \leq k \leq M_{i,i}\} \quad (3)$$

where $M_{i,i}$ is the number of genuine comparisons (i.e., repeated intra-user matchings) for user i. The average scores for user i, namely $\text{avgIntra}(i)$, $\text{avgInterIn}(i)$, and $\text{avgInterOut}(i)$, are computed based on the above-defined sets.

3.2. Inter ZooPlot

ZooPlot defines a scatter plot using $\text{avgIntra}(i)$ as the X-axis and $\text{avgInter}(i)$ as the Y-axis, where each user i is represented as a point based on their scores. Here, the $\text{avgInter}(i)$ for user i is defined as the mean of two separate averages: $\text{avgInterIn}(i)$ and $\text{avgInterOut}(i)$.

Inter ZooPlot uses $\text{avgInterOut}(i)$ as the X-axis and $\text{avgInterIn}(i)$ as the Y-axis to distinguish users based on their behavior when impersonating others and being impersonated. Additionally, $\text{avgIntra}(i)$ is encoded as the color of each user's point, showing the distribution of users across all three score dimensions. Like ZooPlot, Inter ZooPlot belongs to the System-User Level. Figure 2 presents a comparison between the ZooPlot and Inter ZooPlot.

3.3. Biometric Confusion Matrix

a) Standard Confusion Matrix. In non-biometric tasks, confusion matrices are used for evaluation and metric computation. Let's note $\mathcal{F}(x) \rightarrow p$ the classifier $\mathcal{F}$ that classifies a sample $x \in \mathcal{X}$ to a predicted class $p \in C_{train}$. Given a

test dataset $\mathcal{T} = \{(x_i, gt_i) \,|i = (1..N)\}$ of N pairs of sample and groundtruth, the confusion matrix $Conf_F^T$ is a matrix where each row represents a groundtruth class and each column represents a predicted class; each cell $[i, j]$ counts the number of samples of class i predicted as class j. It is built as follows:

$$Conf_F^T[gt, pred] = \sum_{i=1}^{N} \mathbb{1}\left\{y_i = gt \ \wedge \ \mathcal{F}(x_i) = pred\right\}, \quad \forall\ gt \in C_{train},\ pred \in C_{test} \tag{4}$$

Usually, $C_{train} = C_{test}$, but this is not mandatory. Each row sums to the number of samples belonging to the corresponding groundtruth class, and each column to the number of samples predicted as the corresponding class. Notably, each test sample is counted exactly once in the matrix.

Confusion matrices can be challenging to interpret: the cells represent counts (or ratios) for (groundtruth, prediction) pairs, and their interpretation depends on their position in the matrix. Diagonal entries represent correctly classified samples (higher values are better), while off-diagonal entries represent misclassifications (lower values are better).

The confusion matrix is typically visualized as a heatmap, with the color of each cell indicating the number of samples it represents. The same color palette is usually applied across the matrix, including both diagonal and off-diagonal cells. To account for class imbalance, the matrix is often normalized to show ratios instead of raw counts:

$$\tilde{Conf}_F^T[gt, pred] = \frac{Conf_F^T[gt, pred]}{\sum_{i \in C_{test}} Conf_F^T[gt, i]}, \quad \forall\ gt \in C_{train},\ pred \in C_{test} \tag{5}$$

b) Biometric Confusion Matrix (BCM). BCM was designed to better reflect the evaluation needs of BVS, where the process differs from conventional classification. Unlike standard classification, where there is virtually a single reference, in BVS each query in the test set is compared with all biometric references in the training set (the gallery), rather than generating a single prediction. As a result, each query may contribute to multiple entries in the matrix, up to the size of the gallery, rather than being counted once. Let each query x_i be compared against every $u_j \in \mathcal{C}_{\text{train}}$. The function $\mathcal{F}(x_i, u_j) \in \{0, 1\}$ indicates whether the comparison is accepted. The BCM is formally defined as:

$$BCM[gt, pred] = \sum_{i=1}^{N} \sum_{j \in \mathcal{C}_{train}} \mathbb{1}\left\{y_i = gt \wedge \mathcal{F}(x_i, j) = 1 \wedge j = pred\right\}, \quad \forall\ gt \in C_{train},\ pred \in C_{train} \tag{6}$$

This formulation reflects the nature of the evaluation of biometric verification, where decisions are made through many-to-many comparisons rather than single-label predictions. Therefore, careful interpretation of BCM is necessary, as its entries count accepted comparisons, not samples.

To enhance interpretability, the BCM encodes error rates rather than raw counts, promoting a consistent "lower-is-better" reading across all cells. Diagonal elements represent FNMR, while off-diagonal ones correspond to FMR or related quantities. Depending on the evaluation protocol, some entries may be undefined due to limited pairwise comparisons or constraints in test set construction (e.g., partial or sparse verification scenarios). BCMs are typically rendered as heatmaps, with cooler colors for lower errors and warmer tones for higher ones. Distinct color maps can be applied to visually emphasize the difference between FNMR and FMR regions.

c) Users Ordering. BCM is ordered by the following definition. Each row j or column j corresponds to user $order(j)$. The original user ID order from the database doesn't have to be preserved, as it holds no semantic relevance. On the contrary, in the context of Explainable AI, matrix reordering is a form of post-hoc analysis that helps uncover latent structures not explicitly modeled during training. In BCM, such patterns may include user clusters with similar error profiles, asymmetric match tendencies, or outliers. Appropriate sorting strategies enhance visual interpretability, enabling detection of abnormal behaviors or system biases.

Inspired by work on matrix visualization techniques [2] in the broader data analysis field, we aim to adapt similar principles to BCM. Unlike the binary and symmetric matrices typically addressed in prior work, the BCM is real-valued and asymmetric. Therefore, we propose several customized reordering strategies to facilitate effective post-hoc interpretation of BVS performance through BCM.

DATA ordering uses the order of the users as they are represented in the dataset. $order(j) = j$.

FNMR ordering orders users depending on their user FNMR. $order(j) = \text{argsort}\left(\{\text{fnmr}(i) \,|\, i \in \mathcal{U}\}\right)[j]$

FMR In ordering sorts users based on their FMR_{in}. Let $\text{fmr}(i, j)$ denote the FMR when user j (probe) is used to impersonate user i (gallery). $\text{FMR}_{\text{in}}(i)$ is defined as the average over all probes j used to match against i:

$$\text{FMR}_{\text{in}}(i) = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \text{fmr}(i, j) \tag{7}$$

This ordering highlights users who are most frequently impersonated by others: $order(j) = \text{argsort}\left(\{\text{FMR}_{\text{in}}(i) \,|\, i \in \mathcal{U}\}\right)[j]$

FMR Out ordering sorts users depending on their FMR_{out}. Similarly, $\text{FMR}_{\text{out}}(i)$ is defined as the average FMR when user i (probe) is used to impersonate all other users j in the gallery:

$$\text{FMR}_{\text{out}}(i) = \frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \text{fmr}(j, i) \tag{8}$$

This ordering ranks users based on how often they succeed in impersonating others: $order(j) = \text{argsort}\left(\{\text{FMR}_{\text{out}}(i) \mid i \in \mathcal{U}\}\right)[j]$

Average FMR ordering ranks users according to the average of their FMR_{in} and FMR_{out}. This ordering strategy prioritizes users who have higher false match tendencies from both perspectives on average: $order(j) = \text{argsort}\left(\left\{\frac{1}{2}\left(\text{FMR}_{\text{in}}(i) + \text{FMR}_{\text{out}}(i)\right) \,\middle|\, i \in \mathcal{U}\right\}\right)[j]$

VAT ordering aims to reveal latent structures within the BCM by reordering its rows and columns such that similar users are placed close to each other. It is inspired by clustering tendency analysis methods typically applied to pairwise dissimilarity matrices. This ordering relies on the Visual Assessment of Tendency (VAT [14]) algorithm designed to order rows and columns of pairwise dissimilarity matrices. As the BCM doesn't respect such prerequisites, a proxy of pairwise dissimilarity matrix is computed from the BCM, ordered, and uses its order for the BCM.

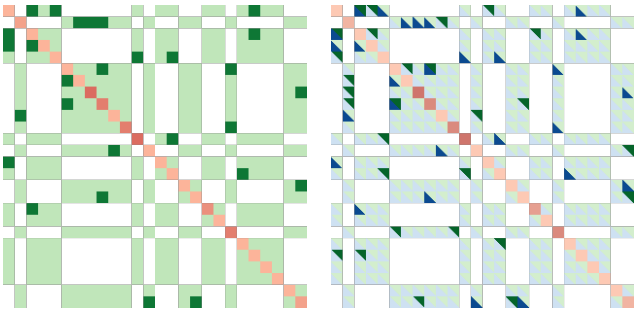
To compute this matrix, the distance between two users is computed by giving the same amount of importance for FMR and FNMR. A vector $repr_i$ of size $2N$ is build for each user i. Each cell is initialized to 0, and the vector is updated as such: $repr_i[i] = repr_i[i+n] = FNMR_i$, $repr_i[j] = FMR_{i \to j}$ and $repr_i[j+n] = FMR_{j \to i}$.

The distance between i and j is computed as such: $distance(i,j) = intra\_error(i,j) + inter\_error(i,j)$ with $intra\_error(i,j) = \sqrt{(repr_i[i] - repr_j[i])^2 + (repr_i[i] - repr_j[i])^2}$ that computes the distance of two users regarding FNMR, and $inter\ error(i,j) = \sqrt{\sum_{k,k \in [1;N] \setminus i,j} (repr_i[k] - repr_j[k])^2}$ that computes the distance of two users regarding all FMR cases.

Let $\text{VAT}(D)$ denote the ordering returned by the VAT algorithm applied to a dissimilarity matrix D. Then, $order(j) = \text{VAT}(D)[j]$.

The matrix reordering feature enables the identification of users with the worst performance in either Inter or Intra Score. Similar to Inter ZooPlot, BCM can distinguish between Lambs and Wolves in the matrix, by user's FMR In/Out. This property supports BCM's capability to categorize all classes within the Biometric Menagerie.

d) Advanced Biometric Confusion Matrix. In the BCM, rows represent FMR_{in} values, while columns represent FMR_{out} values. Naturally, they are not necessarily equal, which makes BCM an asymmetric matrix and inconvenient to directly compare a user's FMR_{in} and FMR_{out} within the matrix. Advanced Biometric Confusion Matrix (Advanced BCM) is proposed to address this limitation. It introduces triangular glyphs, for each off-diagonal cell (i,j), the value displayed in the upper triangle is derived from the symmetric cell (j,i). As a result, asymmetries can be analyzed more directly than in the original BCM, without requiring cross-reference between separate cells.



(a) BCM (b) ABCM

Figure 3. Comparison of BCM and Advanced BCM with missing data (white cells). Advanced BCM improves visual comparison of a user's FMR In/Out, highlighting error symmetry.

Although BCM and Advanced BCM adopt different visualization strategies, they both reveal the distribution of error rates between all users in the system, placing them in the User–User level evaluation category. Figure 3 presents an example of BCM and Advanced BCM.

4. Experiments and Results

Inter ZooPlot and BCM provide a new evaluation perspective on BVS through visualization. We conducted a series of tests to understand how these two visualizations help developers identify user categories and evaluate their explainability capabilities. Section 4.1 provides information about the datasets used in the experiments. Section 4.2 discusses the implementation of the two visualization methods. Section 4.3 investigates the performance of these two visualizations in user category classification through a user study. Section 4.4 discusses the explainability insights derived from their application to real-world datasets.

4.1. Dataset

We used public datasets provided in Table 1. Previous work [6] computed matching score datasets for five biometric datasets: AR [17], ENSIB [7], FC94 [24], FVC [16], and Veins [15] were directly reused. The BSSR1 dataset is from the multimodal face and fingerprint similarity scores dataset released by NIST [20]. It contains similarity scores between users. The BSS1-4 datasets are sourced from the OU-ISIR Biometric Score Database [11, 25]. They contain dissimilarity scores between users, based on gait analysis using videos and sensor-based data. The study uses four different modalities from the database, each involving a different number of users for testing.

Name	Type	Modality	Number of Users	#Inter Scores	#Intra Scores	EER
AR [17]	Similarity Score	Face	120	357,000	3,000	0.102
ENSIB [7]	Similarity Score	Face	100	386,100	3,900	0.109
FC94 [24]	Similarity Score	Face	152	436,088	2,888	0.003
FVC [16]	Similarity Score	Fingerprint	100	69,300	700	0.103
Veins [15]	Similarity Score	Vein	24	16,008	696	0.000
BSSR1_fingface_fc [20]	Similarity Score	Face	517	266,772	517	0.043
BSS_1_GEI [11]	Dissimilarity Score	Gait energy image	2,015(1936)	3,899,104	1,936	0.499
BSS_2_CGI	Dissimilarity Score	Chrono-gait image	3,706	13,730,730	3,706	0.047
BSS_3_Accel [25]	Dissimilarity Score	Gait, accelerometer	737	10,160,300	14,881	0.149
BSS_4_GFI	Dissimilarity Score	Gait flow image	3,249	10,552,752	3,249	0.115

Table 1. Statistics of used datasets. Dataset BSS_1_GEI has a different number of users in Gallery and Probe. Number in parentheses represents the number of probe users. Columns starting with # denote the number of scores.

4.2. Implementation Details

Both visualizations were implemented in Rust for efficient large-scale biometric data processing. The color scales used in the visualizations follow Vega color schemes [22], ensuring perceptual uniformity and clarity. Two rendering approaches were implemented: an interactive interface for detailed user inspection, and a pixel-oriented static visualization optimized for large-scale datasets. The pixel-oriented version was used in the user study.

4.3. User study

a) Experiment Objective. To assess the effectiveness of Inter ZooPlot and BCM in classifying users into the original Biometric Menagerie categories (Lambs, Wolves, Goats, and Sheep), we conducted a user study comparing them with the baseline ZooPlot.

Tested visualizations were divided into two groups. The first group includes the standard versions: *i)* BCM and *ii)* Inter ZooPlot. The second group consists of their variants: *iii)* Inter ZooPlot combined with ZooPlot and *iv)* Advanced BCM. This setup was designed to better reflect realistic usage scenarios: Inter ZooPlot and ZooPlot can be used jointly to provide complementary information, and Advanced BCM can serve as a substitute for BCM depending on developer needs. *v)* ZooPlot is used as the baseline for comparison.

We propose the following hypothesis to compare the classification accuracy of the five visualization methods: ***H1***: Benefiting from the additional information provided, methods *i)–iv)* will achieve higher classification accuracy than the baseline method *v)*.

b) Definition of the GroundTruth. To evaluate the effectiveness of visualizations in our experiments, we required a clear classification criterion as the GroundTruth. Following the approach of Yager and Dunstone [28], we adopted a percentile-based thresholding strategy: the top 75% of users with the most typical scores were labeled as Sheep, while the remaining 25% were categorized as Wolf, Lamb or Goat.

This classification is used only for the purpose of evaluating the discriminative capability of different visualization methods and doesn't imply fixed user categories in real-world systems. The datasets used in the experiments were collected under controlled conditions, whereas real-world biometric verification often takes place in uncontrolled environments. In practice, researchers may adjust the threshold according to their specific application needs.

c) Experiment Design. Twenty-two participants (Age: 27 ± 3, with 15 males) took part in the study. It was conducted in English. Participants had basic knowledge of computer science and statistics.

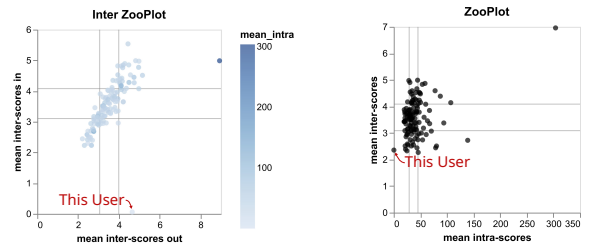
Participants had a 30-minute briefing session covering basic concepts in biometrics, evaluation metrics, user categories, and an introduction to five visualizations they will use before the actual test. After, participants completed a short mock session to confirm their understanding of the task and visualizations. During the formal test, participants were given reference materials containing the content of the briefing session. They then completed 30 user classification tasks involving different systems, users, and methods. Participants' answering times were measured without notifying them during the formal test.

After completing all tasks, participants filled out a 10-point Likert scale questionnaire (1 = very poor, 10 = excellent), rating each visualization in terms of ease of use and perceived certainty of the results. In each case, the participant was informed that the target identity was consistent across all visualizations.

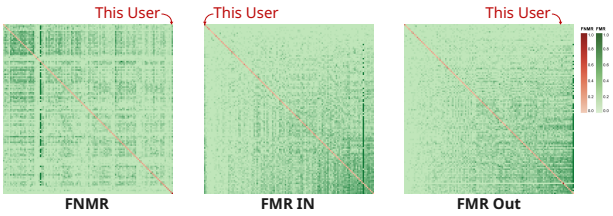
Since BCM and Advanced BCM allow multiple orderings to support classification, their test included three ordering methods: FNMR, FMR In, and FMR Out. These were presented simultaneously. Compared to VAT ordering, which focuses on clustering, they are better suited for classification tasks as they explicitly sort users based on individual error tendencies. To balance the visibility of FNMR and FMR, all BCM and Advanced BCM visualizations were generated using the threshold corresponding to the EER,

Method	Mean Accuracy (All Classes) ↑	Mean Accuracy (Lamb/Wolf) ↑	Mean Usability Rank ↓	Mean Confidence Rank ↓	Mean Answering Time (s) ↓
i) BCM	0.894 (**3**)	0.976 (**1**)	2.20 (**1**)	2.39 (**1**)	14.34 (**1**)
ii) Inter ZooPlot	0.900 (**2**)	0.852 (**4**)	3.59 (**4**)	3.32 (**4**)	27.06 (**4**)
iii) Inter ZooPlot + ZooPlot	0.943 (**1**)	0.919 (**3**)	2.55 (**3**)	2.36 (**2**)	29.02 (**5**)
iv) Advanced BCM	0.875 (**4**)	0.945 (**2**)	2.32 (**2**)	2.39 (**1**)	15.13 (**2**)
v) ZooPlot (Baseline)	0.739 (**5**)	0.631 (**5**)	4.34 (**5**)	4.55 (**5**)	21.65 (**3**)

Table 2. Comparison of methods i) – v). Usability and confidence scores are normalized using Borda count [26]. Arrows indicate direction of desirable values: ↑ means higher is better, ↓ means lower is better. Numbers in parentheses indicate rankings among the five methods for each criterion. Entries with rank 1 are highlighted with a box. All proposed methods outperform the baseline ZooPlot in every criterion except answering time.



(a) Inter ZooPlot + ZooPlot



(b) BCM

Figure 4. Visualizations shown to participants during the test. In the BCM, all three ordering methods were displayed simultaneously. When shown in combination, participants were informed that the visualizations corresponded to the same user.

which ensures comparable error magnitudes between diagonal and off-diagonal regions in the heatmap. Figure 4 shows two examples of user classification tasks using (a) Inter ZooPlot + ZooPlot and (b) BCM.

d) Results Analysis: Table 2 compares the classification accuracy across five visualization methods, shows that the baseline ZooPlot has the lowest accuracy across both metrics, while all other methods showed clear improvements. Accuracy is calculated as the number of correctly identified user categories divided by the total number of actual user categories. The average accuracy per participant was used as the data point. We report both overall classification accuracy and classification accuracy for Lamb/Wolf only (i.e., completely ignore the other classes). To assess statistical significance, Wilcoxon signed-rank tests were performed comparing each method against ZooPlot, with Holm correction applied for multiple comparisons [7, 8]. All comparisons were statistically significant (p-values < 0.05), with mean corrected p-values of 0.010 (all classes) and 0.008 (Lamb/Wolf only), indicating that Inter ZooPlot, BCM, and their variants provide more accurate and reliable support for Lamb/Wolf/Goat/Sheep classification than the baseline. Furthermore, power analysis [3] confirms the robustness of our findings, with mean effect sizes of 0.569 and 0.628 respectively (indicating medium-to-large effects), and corresponding statistical powers of 0.813 and 0.845, which exceed the commonly accepted threshold of 0.8 and thus indicate strong reliability. This supports our hypothesis ***H1***.

Figure 5 shows the distribution of accuracy. Notably, ZooPlot is the only method with a significant drop in Lamb/Wolf accuracy compared to overall accuracy, suggesting its limited ability to distinguish these two classes. While participants could identify when both Lamb and Wolf are present or absent, the aggregated Inter Score made it hard to detect cases where only one is present. This supports the classification limitations of ZooPlot noted by Yager and Dunstone [28].

Table 2 also shows that the combination of Inter ZooPlot and ZooPlot achieved the highest all-class accuracy but also required the longest answering time, which aligns with the intended purpose of the combination—providing more information to support user judgment. Besides, the lower accuracy of Advanced BCM compared to BCM is likely due to the extra information it introduces, which tends to distract participants during classification.

4.4. XAI Potential of the Visualizations

One of the key advantages of using BCM in BVS is its ability to visually reveal error rates between users. By examining the distribution of errors across the matrix, developers can gain insights into the sources of errors for specific users, as well as differences in error distribution across systems. We present two examples to illustrate this point.

Figure 6 shows the position of a specific user in both the Inter ZooPlot and the BCM with ordering FMR Out. In the Inter ZooPlot, this user clearly exhibits high average inter-in and inter-out scores. By examining the BCM, we can un-

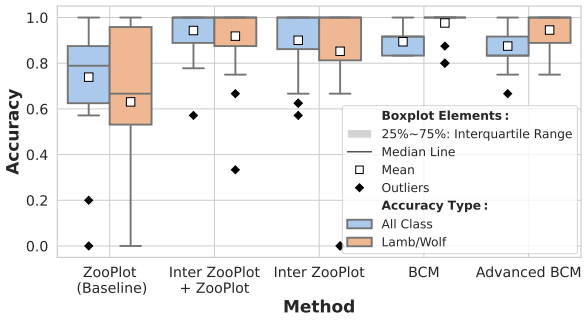


Figure 5. Boxplot of mean accuracy per method. Lamb/Wolf accuracy is computed as the proportion of correctly identified users in each category. In the Lamb/Wolf setting, ZooPlot exhibits lower and more variable performance.

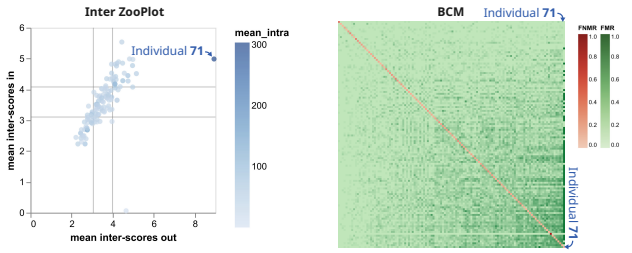


Figure 6. Comparison between the BCM using FMR Out ordering and Inter ZooPlot. BCM provides additional insight into user 71's low performance, indicating that errors are broadly distributed across many users rather than caused by a few outliers.

derstand the underlying reason: this user has a consistently high FMR In when matched against most other users.

Figure 7 compares the BCM visualizations of two datasets with similar EER values, FVC and ENSIB, using a fixed threshold of FMR = 0.01 and VAT ordering to emphasize clustering effects. It can be observed that FVC exhibits a more uniform error distribution, whereas ENSIB shows error concentration around a specific group of users near the corner. This suggests that ENSIB is a less balanced dataset in terms of users' FMR distribution.

These examples highlight BCM's potential as a post-hoc explainability method. By visualizing the user-cluster and distribution of errors, BCM enables developers and researchers to go beyond aggregate metrics and analyze User-User level behaviors and BVS. This capacity to localize and interpret errors aligns with the goals of Explainable AI, providing a human-understandable interface to otherwise opaque system decisions.

5. Conclusion and Future Works

This paper introduced two visualizations, Inter ZooPlot and the Biometric Confusion Matrix (BCM), designed to

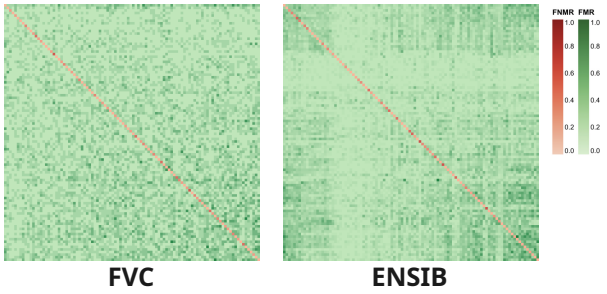


Figure 7. Comparison of BCM using VAT ordering. Despite similar EER values, the ENSIB exhibits visually distinct clustering patterns, suggesting the presence of bias in the BVS.

improve the evaluation and explainability of Biometric Verification System (BVS). They address key limitations of ZooPlot, which struggles to accurately distinguish the original Biometric Menagerie user categories. We also implemented Advanced BCM to improve the visual comparison of user-level error asymmetries. Our user study demonstrated that the proposed visualizations significantly enhance classification accuracy, achieving 90.0% for Inter ZooPlot and 89.4% for BCM compared to 73.9% for ZooPlot. In addition, BCM has also shown potential as a post-hoc explainability method. Its ability to reveal user-specific errors and system-level imbalances supports a deeper understanding of both user behavior and system dynamics. This aligns with the goals of XAI and highlights the importance of interpretable biometric evaluation.

Our future work will focus on three aspects. First, since BCM is implemented as a pixel-oriented visualization currently more suited for small to medium-sized datasets, we plan to explore hierarchical representations and interactive features (e.g., zoom, filtering, group collapsing) to support large-scale use [12]. Second, we aim to integrate additional matrix reordering methods to further enhance the explainability capacity of BCM, helping to expose hidden relational patterns and support a more interpretable and trustworthy evaluation of BVS. Third, a promising direction involves combining BCM and Inter ZooPlot with datasets that include user meta-data, such as gender or ethnicity, to explore potential biases in biometric system performance across different demographic groups. Such integration could offer valuable insights into fairness and accountability in biometric applications.

6. Acknowledgment

This research was funded by the University of Bordeaux and the Scientific Council of LaBRI. We thank the participants in the user study, the biometric database providers, and the anonymous reviewers for their valuable contributions to this work.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. 2

[2] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, and J.-D. Fekete. Matrix reordering methods for table and network visualization. In *Computer Graphics Forum*, volume 35, pages 693–716. Wiley Online Library, 2016. 4

[3] J. COHEN. A power primer. *Psychological bulletin*, 112(1):155–159, 1992. 7

[4] G. R. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds. Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *ICSLP*, volume 98, pages 1351–1354, 1998. 1, 3

[5] T. Dunstone and N. Yager. Biometric system and data analysis: Design, evaluation, and data mining. pages 111–180. Springer, 2009. 1, 2

[6] R. Giot, R. Bourqui, and M. El-Abed. Zoo graph: a new visualisation for biometric system evaluation. In *2016 20th International Conference Information Visualisation (IV)*, pages 190–195. IEEE, 2016. 3, 5

[7] B. Hemery, C. Rosenberger, and H. Laurent. The ensib database: a benchmark for face recognition. In *International Symposium on Signal Processing and its Applications (ISSPA), special session "Performance Evaluation and Benchmarking of Image and Video Processing*, pages 459–464, 2007. 5, 6, 7

[8] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979. 7

[9] R. Ibrahim and M. O. Shafiq. Explainable convolutional neural networks: a taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):1–37, 2023. 1

[10] I. ISO. Iec 19795-1: Information technology–biometric performance testing and reporting-part 1: Principles and framework. *ISO/IEC, Editor*, 1(3):5, 2006. 1, 2

[11] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ouisir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. on Information Forensics and Security*, 7, Issue 5:1511–1521, Oct. 2012. 5, 6

[12] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on visualization and computer graphics*, 6(1):59–78, 2002. 8

[13] A. Krishnan, A. Almadan, and A. Rattani. Investigating fairness of ocular biometrics among young, middle-aged, and older adults. In *2021 International Carnahan Conference on Security Technology (ICCST)*, pages 1–7. IEEE, 2021. 2

[14] D. Kumar and J. C. Bezdek. Visual approaches for exploratory data analysis: A survey of the visual assessment of clustering tendency (vat) family of algorithms. *IEEE Systems, Man, and Cybernetics Magazine*, 6(2):10–48, 2020. 5

[15] P.-O. Ladoux, C. Rosenberger, and B. Dorizzi. Palm vein verification system based on sift matching. In *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*, pages 1290–1298. Springer, 2009. 5, 6

[16] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *2002 International conference on pattern recognition*, volume 3, pages 811–814. IEEE, 2002. 5, 6

[17] A. Martinez and R. Benavente. The ar face database: Cvc technical report, 24. 1998. 5, 6

[18] J. R. Matey, G. W. Quinn, P. Grother, E. Tabassi, C. Watson, and J. L. Wayman. Modest proposals for improving biometric recognition papers. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2015. 2

[19] T. J. Neal and D. L. Woodard. Surveying biometric authentication for mobile device security. *Journal of Pattern Recognition Research*, 1(74-110):4, 2016. 1

[20] N. I. of Standards and Technology. Nist biometric scores set (bssr1), 2017. 5, 6

[21] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. An introduction evaluating biometric systems. *Computer*, 33(2):56–63, 2000. 1

[22] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE transactions on visualization and computer graphics*, 22(1):659–668, 2015. 6

[23] M. E. Schuckers. *Computational methods in biometric authentication: statistical methods for performance evaluation*. Springer Science & Business Media, 2010. 2

[24] L. Spacek. Faces94 face image dataset. https://csperson.kku.ac.th/chakchai/faceDBweb/face_dataset.html, 1994. Accessed: 2025-04-10. 5, 6

[25] N. Trung, Y. Makihara, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi. Performance evaluation of gait recognition using the largest inertial sensor-based gait database. In *The 5th IAPR International Conference on Biometrics (ICB 2012), Mar 29 - Apr 1, New Delhi*, Mar 2012. 5, 6

[26] M. Van Erp and L. Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *7th International Workshop on frontiers in handwriting recognition*, pages 443–452. International Unipen Foundation, 2000. 7

[27] J. Wayman, A. Jain, D. Maltoni, and D. Maio. An introduction to biometric authentication systems. In *Biometric systems: Technology, design and performance evaluation*, pages 1–20. Springer, 2005. 1

[28] N. Yager and T. Dunstone. The biometric menagerie. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):220–230, 2008. 1, 3, 6, 7