Novelty Detection in Human-Machine Interaction Through a Multimodal Approach

José Salas-Cáceres[1 [0009−0004−7543−3385]], Javier Lorenzo-Navarro[1 [0000−0002−2834−2067]], David Freire-Obregón[1 [0000−0003−2378−4277]], and Modesto Castrillón-Santana[1 [0000−0002−8673−2725]]

Universidad de Las Palmas de Gran Canaria
Instituto Universitario SIANI, Las Palmas, SPAIN
jose.salas@ulpgc.es

Abstract. As the interest in robots continues to grow across various domains, including healthcare, construction and education, it becomes crucial to prioritize improving user experience and fostering seamless interaction. These human-machine interactions (HMI) are often impersonal. Our proposal, built upon previous work in the field, aims to use biometric data of individuals to detect whether a person has been encountered before. Since many models depend on a threshold set, an optimization method using a genetic algorithm was proposed. The novelty detection is made through a multimodal approach using both voice and facial images from the individuals, although the unimodal approaches of just each single cue were also tested. To assess the effectiveness of the proposed system, we conducted comprehensive experiments on three diverse datasets, namely VoxCeleb, Mobio and AveRobot, each possessing distinct characteristics and complexities. By examining the impact of data quality on model performance, we gained valuable insights into the effectiveness of the proposed solution. Our approach outperformed several conventional novelty detection methods, yielding superior and therefore promising results.

Keywords: Novelty Detection · Human-Machine Interaction · Biometrics

1 Introduction

The interest in robots continues to rise over the years [19], and this growing fascination is well-founded. These machines have demonstrated a multitude of applications in various domains such as healthcare [21], construction [22], among others. Consequently, there is an increasing number of human-machine interactions (HMI) involving what are known as social robots [24]. Social robots are specifically designed to interact with humans and typically assist them in different tasks. Enhancing the user experience poses a challenge in creating more natural and personal interactions in this scenario. It has been observed that

Table 1: Modified extract from a table obtained from [5]

Task	Training Classes	Test Classes	Objective
Traditional Classifier	KKCs	KKCs	Classify data into one of the known classes.
Reject Option Classifier	KKCs	KKCs	Classify data and reject samples with low confidence.
Outlier Detection	KKCs and some KUCs samples	KKCs and KUCs	Detect outliers in the data.
Novelty Detection	KKCs	KKCs and UUCs	Differentiate between UUCs and KKCs.
Open-Set Classifier	KKCs	KKCs and UUCs	Identify samples belonging to known classes and categorize them correctly if they do belong.

people respond better to HMI if they are recognized by the robot, only if the interaction is after a previous encounter. Therefore, this work aims to develop a HMI model capable of detecting whether a person has been encountered before. This model would utilize biometric data of the individuals. By doing so, if these individuals run into the same robot again, the model would allow the robot to recognize them. To summarize, the goal is to design a novelty detection model for individuals based on biometrics and with the capability of efficiently and quickly expanding the database of enrolled identities.

2 Related work

2.1 Terminology

Several terms will be used throughout this work in the context of novelty detection. First, we introduce the concept of Out of Distribution (OOD) data, which refers to data encountered during model exploitation that were not present in the training set. There are two types depending on their relationship with the original domain: novelties and anomalies. Novelties are related to the working domain, while anomalies are not. Next, a classification of the different types of classes based on their appearance in the training set and the knowledge about them is made, resulting in four categories [5, 14]:

- **Known Known Class** (KKC): Refers to classes belonging to the known categories used to train the model.
- **Known Unknown Class** (KUC): Represent classes belonging to a class not in the KKCs but represented in the training set.
- **Unknown Unknown Class** (UUC): Denotes classes that belong to unknown categories and are not encountered during the training phase.
- **Unknown Known Class** (UKC): Indicates classes belonging to known categories but with no specific samples in the training set; instead, only another type of information is known.

Based on this classification, several tasks arise, presented in Table 1. Among these tasks, this work focuses on novelty detection. To achieve this, a multimodal

approach will be employed, combining data to enhance the performance of the models. Specifically, facial images and voice recordings are going to be used. Due to the transient nature of the interaction, the amount of data taken from each person will be limited. As we said before, the database of enrolled persons has to be able to expand continuously. Therefore, selecting models poses a challenge, as some methods require a long time to train with the new data. This limitation will exclude models based on deep learning and neural networks with high training computational demands.

2.2 Existing modeling architectures

In the literature, multiple models have been proposed to address the task of novelty detection. Some will be mentioned here, especially those considered the most suitable for the selected scenario.

First, density-based models are considered, examining the spatial density across different regions to discern whether a sample is a novelty. Among these models, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [3] stands out as a clustering algorithm that groups samples based on proximity and classifies those located further of a certain distance as noise or, in our case, as a novelty. Another noteworthy approach is the Local Outlier Factor (LOF) [1], which leverages the notion of local density computed via distances to the K nearest neighbors. Shifting to classifier-based methods, various one-class classifiers are used in the literature. Below we employ the One-Class Support Vector Machine (OCSVM) [16]. Like other SVM-based models, this algorithm aims to maximize the margin between samples of distinct classes by using a hyperplane, being the one class of the binary classifier, identities in the dataset, ergo not novelties. Furthermore, Support Vector Data Description (SVDD) [20] has a similar principle but employs a hypersphere to enclose the classes instead of a hyperplane. Additionally, Isolation Forest (IF) [8] is often used in novelty detection. This algorithm, rooted in decision trees, strives to isolate samples. The underlying concept is that if a sample is quickly isolated, it is likely to be an outlier; conversely, if it is challenging to segregate from the rest, it is not an outlier. Lastly, we consider two models based on probability. Gaussian Mixture Model (GMM) fits Gaussian distributions to the available data. Alternatively, Kernel Density Estimation (KDE) [6] estimates the density of a given set of points by aggregating different kernels, such as Gaussian or exponential distributions. Then, these probability-based models compute a probability of a sample belonging to the distribution made, setting a threshold is possible to differentiate between regular new samples and OOD data. There are other methods for novelty detection beyond those mentioned, including those based on reconstruction or deep learning [15, 25]. It is also important to assert that many algorithms depend on setting up a threshold, limiting the number of ramifications for IF or the probability of belonging to a specific distribution like in KDE.

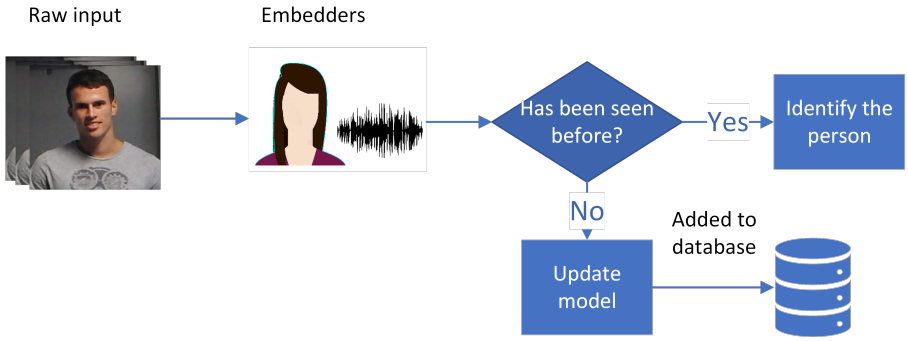

Fig. 1: Open-set classifier scheme.

3 Methodology

Figure 1 depicts the intended behavior of the proposal. Upon detecting a person, the robot captures his/her voice and face. This raw data then undergoes a preprocessing stage, being transformed into feature vectors. These vectors are then fed into trained models to determine if the person is known or unknown. The human-machine interface (HMI) proceeds uninterrupted if the person is recognized. However, if the person identity is unknown, the robot updates the database with the new identity.

As it was mentioned above, in our experiments we did not directly process the raw audio or image data. Instead, we employed a preprocessing step to convert the samples into fixed-dimensional numerical vectors known as embeddings. These embeddings were generated using specific neural networks called embedders, which were trained specifically for this task.

- For voice samples, we utilized the **X-Vector** network [18], which is trained to discriminate between different speakers. It was designed to convert audio of variable duration into fixed-dimensional vectors.
- For facial images, we employed the **FaceNet** network [17], which is trained to map facial images to a Euclidean space where distances reflect facial similarity. Similar to X-Vector, FaceNet generates fixed-dimensional embeddings.

By employing these dedicated embedders, we were able to transform the raw voice and facial image data into standardized and informative numerical representations. The two embedders used in our approach generate vectors of 512 elements each. In the multimodal approach, we concatenate these voice and face embeddings, resulting in a final feature vector of 1024 dimensions.

For our practical experiments, we selected a subset of those models described above in the related work section and applied them to our specific scenario. Those chosen models served as the foundation for our evaluation and analysis.

In our context, a person belonging to a KKC is someone who is already in the database. One in the UUC represents one that is not, a novelty. There is not

KUC, however, that would encompass individuals registered in the database as a unknown.

4 Datasets

As previously mentioned, a multimodal approach will be adopted, requiring audiovisual data. Three distinct available audiovisual datasets were utilized, each with its characteristics and complexities.

First, let us describe AveRobot [9], a dataset specifically created with HMI in mind. The dataset consists of approximately 10-second videos where different individuals simulate interactions with a robot. These videos were recorded with eight different sensors in several indoor locations of a realistic and everyday environment, precisely the common spaces of a university building. Given those real life characteristics, the illumination conditions in these locations were not optimal, resulting in poor image quality, including noise, blurriness, and lighting issues. The audio quality suffers from a similar condition. The AveRobot dataset comprises samples from 111 individuals, most falling within 15 to 25 years. This dataset has been successfully used for multimodal user verification [4]. Another audiovisual dataset evaluated is VoxCeleb 1 [13], which consists of interviews with celebrities posted on Youtube. This dataset contains samples from 1251 celebrities from around the world. While the dataset exhibits large diversity, there is a predominant representation of males and native English speakers. Furthermore, owing to the data extraction source, the image and audio quality in VoxCeleb 1 are exceptionally high. Because of this, this dataset may not entirely represent the data one would encounter when attempting to integrate a model into a HMI environment. Lastly, the audiovisual dataset Mobio [7] was studied. This dataset was recorded using two mobile devices: one being a Nokia N93i mobile phone and the other being a standard 2008 MacBook laptop. The dataset consists of over 61 hours of audiovisual data with 12 distinct sessions usually separated by several weeks. In total there are 192 unique audiovideo samples for each participant. This data was captured at 6 different sites over one and a half years with people speaking English. In this paper, we used the training and evaluation partitions, this two subsets have a total of 92 identities. The distinction in quality between the three datasets can be seen in Figure 2.

5 Experiments

All experiments in our study followed a standardized structure, consisting of the following steps:

1. **Data loading**: Prior to conducting the experiments, three subsets of data were generated for each dataset described in Section 4: training, validation, and testing. The specific characteristics of each subset can be found in Table 2. It should be noted that for the experiments conducted with the Mobio

(a) AveRobot samples [9]. **(b)** VoxCeleb samples [13]. **(c)** Mobio samples [7].

Fig. 2: Example images of the datasets.

Table 2: Characteristics of Data Subsets. Ids stands for identities.

Set	# known ids	# unknown ids	# images per id	# audios per id
Train	50	0	50	2
Validation	50	50	30	1
Test	50	50	30	1

dataset, 46 individuals were used per set, and 40 images were used per sample in the training set. This was due to limitations in the number of identities available in it.

2. **Model training**: After loading the data, each evaluated model was trained using the training set. Hyperparameter tuning was performed using a grid search with different combinations and leveraging the validation set. In some cases, we also tested different methods for calculating the threshold. It is important to note that during the training phase, there were no unknown samples. For threshold calculation, only elements from the training set, which all belonged to the Known Known Classes (KKCs), were used.
3. **Performance testing**: The model's performance was evaluated using the test set. Two separate tests were conducted: one to assess the model's ability to detect known samples and another to evaluate its capability in detecting novel samples. The final results were derived from a combination of the model's performance in both tests.

It is worth mentioning that for most models, we conducted a small search to identify the best possible hyperparameters. Furthermore, each model was tested using each dataset. In the following subsections, we provide a brief overview of the different experiments conducted in our study.

5.1 Distance-based experiment

The first experiment conducted is a variant of the Nearest Class Mean (NCM) algorithm [12], a classification algorithm that calculates a centroid for each class,

and then the distances between each new sample and all the centroids are computed, and the sample is assigned to the class whose have the nearest centroid. This work has designed a modified version to adapt NCM for novelty detection.

The proposed experiment has k classes, where k is determined by the number of known individuals in the database at a specific moment. Each class is defined by n_i samples and a centroid $\bar{X}_i$, calculated using the formula 1.

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{ni} \boldsymbol{X}_{ij} \tag{1}$$

When considering a new sample $\boldsymbol{X}_{new}$, it can either belong to one of the KKCs or be a novelty. This is determined by a threshold Thr_i, which can be calculated in various ways, always as a linear combination of some of the distances presented in the formulas: 2, 3, 4, 5, 6 and 7:

$$\boldsymbol{d}_m = \frac{1}{n_i} \sum_{j=1}^{ni} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2 \tag{2}$$

$$\boldsymbol{d}_M = \max_{j=1,2,..,n_i} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2 \tag{3}$$

$$\boldsymbol{D}_m = \frac{1}{k} \sum_{i=1}^{k} (\frac{1}{n_i} \sum_{j=1}^{ni} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2) \tag{4}$$

$$\boldsymbol{D}_M = \max_{i=1,2,..,k} (\max_{j=1,2,..,n_i} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2) \tag{5}$$

$$\boldsymbol{D}_{mM} = \frac{1}{k} \sum_{i=1}^{k} (\max_{j=1,2,..,n_i} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2) \tag{6}$$

$$\boldsymbol{D}_{Mm} = \max_{i=1,2,..,k} (\frac{1}{n_i} \sum_{j=1}^{ni} ||\boldsymbol{X}_{ij}, \bar{X}_i||_2) \tag{7}$$

Thr_i forms a radius around $\bar{X}_i$, as shown in Figure 3. According to expression 8, if the new sample falls within the influence zone of $\bar{X}_i$ it will be considered part of one of the known classes (KKC). If not, it will be considered a novelty (UUC).

$$\begin{cases} \boldsymbol{X}_{new} \in \text{KKC}, \text{ if } \exists i \text{ s.t. } ||\boldsymbol{X}_{new}, \bar{X}_i||_2 <= Thr_i \\ \boldsymbol{X}_{new} \notin \text{KKC}, \text{ if } ||\boldsymbol{X}_{new}, \bar{X}_i||_2 > Thr_i \forall i \end{cases} \tag{8}$$

This experiment was conducted in five different cases, which differed in the type of data used, the centroid calculation process, or in the application of an additional step:

– **Unimodal**: To have a basis for comparison with the multimodal approach, the generated feature vectors from voice and images were separately used for analysis.

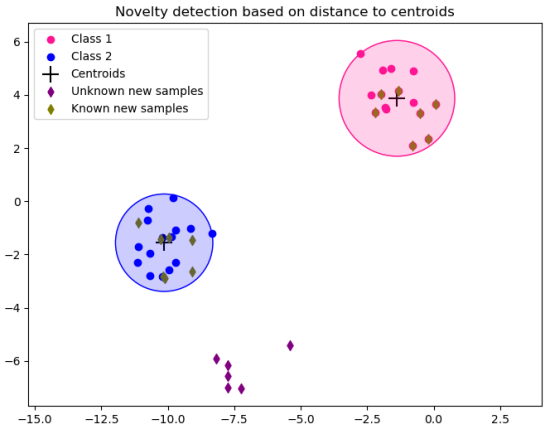

Fig. 3: Visual representation of the explained process for the NCM variant. The crosses depict the centroids, the diamonds represent the new samples, and the dots represent the existing samples in the databases. It is observed how the radii delimit the space that belongs to each class.

- **Multimodal**: The feature vectors from the unimodal cases were concatenated, forming a vector with twice the dimension of the original embeddings.
- **Multimodal variants**: Two variants were explored. One applies a dimensionality reduction technique, such as PCA, to the concatenated vectors. The second variant utilized GMM to calculate multiple centroids per class instead of a single centroid. These centroids corresponded to the mean positions of the Gaussian distributions that the GMM fitted to the data of each individual. Principal Component Analysis (PCA) was applied to achieve a 95% level of representation, resulting in a reduced dimensionality of 85 elements.

Various strategies for the Thr calculation were tested to find the one that gives the best results in each case. The different strategies are represented in Table 3, each σ the result of the formulas 9, 10 and 11.

$$\sigma_d = \sqrt{\frac{\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{X}_i)^2}{n_i - 1}} \tag{9}$$

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(\boldsymbol{d}_m - \bar{d}_m)^2}{k - 1}} \tag{10}$$

$$\sigma_M = \sqrt{\frac{\sum_{i=1}^{k}(\boldsymbol{d}_M - \bar{d}_M)^2}{k - 1}} \tag{11}$$

The results and the best threshold strategy for each case can be found in Table 4. The results obtained on the VoxCeleb 1 dataset are visibly better than

Table 3: Different methods used to calculate the threshold in the experiments

Threshold strategy	Equation
T_0	$Thr = 3 * \sigma_d$
T_1	$Thr = d_m + 3 * \sigma_d$
T_2	$Thr = d_M$
T_3	$Thr = d_M - 3 * \sigma_d$
T_4	$Thr = D_M$
T_5	$Thr = D_M + 3 * \sigma_M$
T_6	$Thr = D_M - \sigma_M$
T_7	$Thr = D_M - 3 * \sigma_d$
T_8	$Thr = D_{mM} + 3 * \sigma_M$
T_9	$Thr = D_{mM} + 3 * d_m$
T_{10}	$Thr = D_{Mm} + \sigma_m$
T_{11}	$Thr = D_{Mm} + 3 * \sigma_m$

Table 4: Results obtained for the three datasets in each one of the cases. Highlighted values correspond to best results for each set.

	AveRobot			**VoxCeleb 1**			**Mobio**		
Experiment	F1	Acc.	Thr	F1	Acc.	Thr	F1	Acc.	Thr
Face	0.8857	0.8953	T_3	0.9837	0.9837	T_2	0.9725	0.9728	T_3
Voice	0.0000	0.5000	T_0	0.5915	0.7100	T_5	0.9200	0.9130	T_5
Multimodal	0.8636	0.8557	T_{10}	**0.9977**	**0.9977**	T_{10}	0.9683	0.9674	T_6
Multi. PCA	**0.8982**	**0.8967**	T_5	0.9910	0.9910	T_1	0.9831	0.9833	T_9
Multi. GMM	0.3931	0.6223	T_0	0.9934	0.9933	T_5	**0.9856**	**0.9855**	T_5

those achieved with AveRobot, the same happens with Mobio. This can be attributed to the noise and data conditions, which pose a more significant challenge for the models in the AveRobot dataset. It is also apparent that the best results for each dataset are obtained from multimodal approaches. However, it is worth noting that the unimodal option that uses the subject's face also yields good results and that, in Mobio, use the voice alone lead to much better results that in the other datasets.

5.2 Distribution and density-based experiments

The second experiment involved applying the previously explained KDE (Kernel Density Estimation) algorithm. The novelty detection in this algorithm is also based on setting a threshold, in this case, on the score contributed by the model indicating the likelihood of a sample belonging to the distribution constructed with the train data. The threshold Thr will be the same for all the samples and will be calculated so that all training samples are always considered known. To achieve this, Thr is set at the 0th and 100th percentiles, ensuring that any new sample obtaining a score outside the original distribution of scores will be

Table 5: Results obtained using KDE and HDBSCAN for the three datasets.

	AveRobot		**VoxCeleb 1**		**Mobio**	
Experiment	F1	Acc.	F1	Acc.	F1	Acc.
KDE	0.5811	0.6333	**0.9488**	**0.9513**	**0.9528**	**0.9522**
HDBSCAN	**0.7004**	**0.6703**	0.9105	0.9167	0.9094	0.9025

considered a novelty. The aforementioned score is calculated as the logarithm of the estimated density.

The density-based experiment utilizes a variant of the DBSCAN algorithm called Hierarchical DBSCAN (HDBSCAN) [2]. HDBSCAN applies the original DBSCAN algorithm with different radius values and integrates the results to find the most stable clustering [10]. The implementation used [11] can generate a score representing the probability of a new sample being OOD. This score is calculated using the GLOSH algorithm, a variant of the mentioned LOF that compares the density of the space where a sample is located with the density of the samples associated with it [2]. Similar to KDE, novelty detection sets a threshold Thr between the 0th and 100th percentiles.

The results obtained using KDE and HDBSCAN can be found in Table 5. Similar to the previous case, the performance achieved with VoxCeleb 1 and Mobio are significantly better than those achieved with AveRobot. Its worth noting that KDE performs better than HDBSCAN in those dataset with better sample quality (Mobio and VoxCeleb 1) but in AveRobot HDBSCAN is more effective, this is because HDBSCAN is designed to exhibit more robust behavior in noisy situations compared to KDE, which is more sensitive to the noise.

5.3 Non threshold-based models

In addition to the threshold-based models mentioned above, we also evaluated some classification-based models that do not require any adjustment for threshold calculation. They rely on training with the available data. The results can be seen in Table 6. We can notice that the results are much better in VoxCeleb 1 and Mobio. Additionally, these models do not adapt well to the specific problem a hand, as in almost all cases, accuracy higher than 60% is not achieved. An exception to this is seen in the OCSVM though, which achieve an accuracy above 70% in VoxCeleb 1 and Mobio. Another notable point is the substantial difference between the F1-Score and accuracy in some cases, such as KNN. This is due to a high disparity between Recall and Precision, indicating that either the model classified almost all new samples as known or all samples were considered novelties.

6 Performance Optimization

In order to improve the performance of models that rely on a threshold value for determining novelty or known samples, the genetic algorithm (GA) was utilized,

Table 6: Results using a variety of models in each dataset. Highlighted values correspond to best results.

	AveRobot		**VoxCeleb 1**		**Mobio**	
Model	F1	Acc.	F1	Acc.	F1	Acc.
OCSVM	0.6028	0.4773	**0.7540**	**0.7043**	**0.7816**	**0.7435**
LOF	0.5135	0.5207	0.2100	0.5587	0.4108	0.6290
IF	**0.6633**	**0.5043**	0.6709	0.5053	0.7475	0.6736
SVDD	0.6046	0.5100	0.5798	0.5077	0.6537	0.5540
KNN	0.1550	0.5420	0.1355	0.5363	0.2330	0.5659

which is a bio-inspired heuristic optimization method [23]. A single-objective approach was adopted, where the accuracy of the models was maximized. To achieve this, the chromosome was encoded to explore various methods of calculating the threshold for each case and test previously unexplored combinations of hyperparameters. The threshold optimization was performed using train and validation subsets. Once the final performance was obtained, the configuration that yielded to the best results was tested on the validation set, resulting in the values presented in Table 8.

6.1 NCM-based algorithm

In this case, the chromosome consisted of six genes $[G_1, G_2, G_3, G_4, G_5, G_6]$, each limited to vary within the range of -5.0 to 5.0, being always a rational number. From these genes, Thr_i was calculated using the expression described in equation 12.

$$Th_i = G_1 * \boldsymbol{d}_{mi} + G_2 * \sigma_d + G_3 * \boldsymbol{D}_{mM} + G_4 * \sigma_M + G_5 * \boldsymbol{D}_m + G_6 * \sigma_m \quad (12)$$

6.2 Distribution and density-based approaches

For both models, the chromosome structure follows the same pattern, four gens $[G_1, G_2, G_3, G_4]$ where two are used for hyperparameter exploration and the other two for Thr calculation. The structure is shown in Table 7. The objective of this organization is twofold: firstly, to explore different configurations of hyperparameters and secondly, to vary the threshold location, all to improve the performance. The threshold will be calculated as indicated in eq. 13, where L represents the limits obtained from the G_3 percentile and the $(100 - G_3)$ percentile and σ_{scores} is the standard deviation of the scores obtained from the training set.

$$Th = L_{G_3} \pm G_4 * \sigma_{scores} \quad (13)$$

In Table 8, a comparison of the results obtained by applying GA concerning the initial results is presented. Overall, there is an improvement, whether more or less significant, in the performance. Although all the results may not

Table 7: Codification of the chromosomes used in the GA for KDE and HDBSCAN.

	KDE	
Gen	Hyperparameter	Search Space
G_1	kernels	$0 < G_1 < 3. \quad G_1 \in \mathbb{N}$
G_2	bandwidth	$0 < G_2 < 1. \quad G_2 \in \mathbb{R}$
G_3	Percentile	$95 < G_3 < 100. \quad G_3 \in \mathbb{N}$
G_4	σ_{scores}	$-5 < G_4 < 1. \quad G_2 \in \mathbb{R}$
	HDBSCAN	
Gen	Hyperparameter	Search Space
G_1	min_cluster_size: 10	$2 < G_1 < 15. \quad G_1 \in \mathbb{N}$
G_2	min_samples	$2 < G_2 < 15. \quad G_2 \in \mathbb{N}$
G_3	Percentile	$95 < G_3 < 100. \quad G_3 \in \mathbb{N}$
G_4	σ_{scores}	$-1 < G_4 < 5. \quad G_2 \in \mathbb{R}$

improve significantly, it is important to note that the thresholds shown in Table 3 were obtained through trial and error, requiring multiple attempts and with no theoretical base. In contrast, using the GA to calculate these values requires minimal human intervention. Additionally, in the results of the experiment using only voice, a significant improvement is observed. This is likely because a good expression was not found in the trial-and-error process for setting the threshold. The same applies to the multimodal application of GMM in AveRobot.

7 Conclusions

Various techniques for novelty detection of individuals based on their biometrics have been developed throughout this work. The NCM-based algorithm has demonstrated the best performance in every dataset, specifically its multimodal application combining facial image and voice data. Furthermore, due to the high dimensionality of the problem, applying dimensionality reduction techniques such as PCA has been shown to decrease complexity without sacrificing performance.

Another essential aspect observed during the study is that threshold-based models, such as the mentioned implementation of NCM or KDE, are highly dependent on the proper adjustment of their hyperparameters. Therefore, it is considered good practice to use optimization methods to explore various combinations to find a practical expression. This is particularly important when considering that the optimal strategy for threshold calculation not only varies with the nature of the data but also with the dataset employed.

It is worth mentioning that the conditions under which the data is collected significantly impact the performance. This is evident in the apparent differences in results between the datasets used. Finally, distance-based or density-based models have been deemed the best option due to the data limitation and the desired training agility.

Table 8: The results obtained by applying the GA. In each of the metrics △ represent the difference with the result achieved in the previous experiment.

	AveRobot			
Experiment	F1-Score	△F1-Score	Accuracy	△Accuracy
NCM. Face	0.889	0.0033	0.8953	0
NCM. Voice	0.6095	0.6095	0.59	0.09
NCM. Multimodal	0.8966	0.033	0.8997	0.044
NCM. Multimodal. PCA	0.9046	0.0064	0.911	0.0143
NCM. Multimodal. GMM	0.8228	0.4297	0.8393	0.217
KDE	0.724	0.1429	0.723	0.0897
HDBSCAN	0.6983	-0.0021	0.6923	0.022
	VoxCeleb 1			
Experiment	F1-Score	△F1-Score	Accuracy	△Accuracy
NCM. Face	0.9891	0.0054	0.989	0.0053
NCM. Voice	0.9184	0.3269	0.92	0.21
NCM. Multimodal	0.9973	-0.0004	0.9973	-0.0004
NCM. Multimodal. PCA	0.997	0.006	0.997	0.006
NCM. Multimodal. GMM	0.9973	0.0039	0.9973	0.004
KDE	0.9902	0.0414	0.9903	0.039
HDBSCAN	0.9065	-0.004	0.912	-0.0047
	Mobio			
Experiment	F1-Score	△F1-Score	Accuracy	△Accuracy
NCM. Face	0.9803	0.0078	0.9804	0.0076
NCM. Voice	0.9247	0.0047	0.9239	0.0109
NCM. Multimodal	0.9942	0.0259	0.9942	0.0268
NCM. Multimodal. PCA	0.9772	-0.0059	0.9772	-0.0061
NCM. Multimodal. GMM	0.9902	0.0046	0.9902	0.0047
KDE	0.988	0.0352	0.988	0.0358
HDBSCAN	0.9273	0.0179	0.925	0.0225

The logical next step following this work is to develop an Open-Set classifier that can not only successfully perform novelty detection but also classify the identified samples into their respective KKC. Additionally, despite the emphasis on agility mentioned earlier, exploring solutions based on deep learning, particularly those that fall under the Few-Shot learning paradigm, which requires only a small amount of training data, would be interesting.

8 Acknowledgments

This work is partially funded by the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22 and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04, it is also supported by "Programa Investigo" refference code 32/39/2022-0923131539 of Servicio Canario de Empleo. "Fondos del Plan de Recuperación, Transformación y Resiliencia - Next Generation EU".

References

1. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (may 2000). https://doi.org/10.1145/335191.335388, https://doi.org/10.1145/335191.335388
2. Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J.: Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans. Knowl. Discov. Data **10**(1) (jul 2015). https://doi.org/10.1145/2733381, https://doi.org/10.1145/2733381
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
4. Freire-Obregón, D., Rosales-Santana, K., Marín-Reyes, P.A., Penate-Sanchez, A., Lorenzo-Navarro, J., Castrillón-Santana, M.: Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. Pattern Recognition Letters **149**, 179–184 (2021). https://doi.org/10.1016/j.patrec.2021.06.014
5. Geng, C., Huang, S., Chen, S.: Recent advances in open set recognition: A survey. CoRR **abs/1811.08581** (2018), http://arxiv.org/abs/1811.08581
6. Hu, W., Gao, J., Li, B., Wu, O., Du, J., Maybank, S.: Anomaly detection using local kernel density estimation and context-based regression. IEEE Trans. on Knowl. and Data Eng. **32**(2), 218–233 (feb 2020). https://doi.org/10.1109/TKDE.2018.2882404, https://doi.org/10.1109/TKDE.2018.2882404
7. Khoury, E., El Shafey, L., McCool, C., Günther, M., Marcel, S.: Bi-modal biometric authentication on mobile phones in challenging conditions. Image and Vision Computing pp. 1147–1160 (2014). https://doi.org/http://dx.doi.org/10.1016/j.imavis.2013.10.001, http://www.sciencedirect.com/science/article/pii/S0262885613001492
8. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17
9. Marras, M., Marín-Reyes, P.A., Navarro, J.J.L., Santana, M.F.C., Fenu, G.: Averobot: an audio-visual dataset for people re-identification and verification in human-robot interaction. ICPRAM (Setúbal) (2019). https://doi.org/10.5220/0007690902550265
10. McInnes, L., Healy, J.: Accelerated hierarchical density based clustering. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE (nov 2017). https://doi.org/10.1109/icdmw.2017.12, https://doi.org/10.1109%2Ficdmw.2017.12
11. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. The Journal of Open Source Software **2**(11), 205 (2017)
12. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(11), 2624–2637 (2013). https://doi.org/10.1109/TPAMI.2013.83
13. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)

14. Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., Sabokrou, M.: A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. CoRR **abs/2110.14051** (2021), https://arxiv.org/abs/2110.14051
15. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. CoRR **abs/1703.05921** (2017), http://arxiv.org/abs/1703.05921
16. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Solla, S., Leen, T., Müller, K. (eds.) Advances in Neural Information Processing Systems. vol. 12. MIT Press (1999)
17. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
18. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333 (2018). https://doi.org/10.1109/ICASSP.2018.8461375
19. Stock-Homburg, R.: Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. International Journal of Social Robotics **14**(2), 389–411 (Mar 2022). https://doi.org/10.1007/s12369-021-00778-6, https://doi.org/10.1007/s12369-021-00778-6
20. Tax, D.M., Duin, R.P.: Support vector data description. Machine Learning **54**(1), 45–66 (Jan 2004). https://doi.org/10.1023/B:MACH.0000008084.60811.49, https://doi.org/10.1023/B:MACH.0000008084.60811.49
21. Uluer, P., Kose, H., Gumuslu, E., Barkana, D.E.: Experience with an affective robot assistant for children with hearing disabilities. International Journal of Social Robotics **15**(4), 643–660 (Apr 2023). https://doi.org/10.1007/s12369-021-00830-5, https://doi.org/10.1007/s12369-021-00830-5
22. Wang, X., Liang, C.J., Menassa, C.C., Kamat, V.R.: Interactive and immersive process-level digital twin for collaborative human–robot construction work. Journal of Computing in Civil Engineering **35**(6), 04021023 (2021). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000988, https://ascelibrary.org/doi/abs/10.1061/(ASCE)CP.1943-5487.0000988
23. Whitley, D.: A genetic algorithm tutorial. Statistics and Computing **4**(2), 65–85 (Jun 1994). https://doi.org/10.1007/BF00175354, https://doi.org/10.1007/BF00175354
24. Youssef, K., Said, S., Alkork, S., Beyrouthy, T.: A survey on recent advances in social robotics. Robotics **11**(4) (2022). https://doi.org/10.3390/robotics11040075, https://www.mdpi.com/2218-6581/11/4/75
25. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 665–674. KDD '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098052, https://doi.org/10.1145/3097983.3098052