applied sciences

MDPI

Article

Generalized Replay Spoofing Countermeasure Based on Combining Local Subclassification Models

Sarah Mohammed Altuwayjiri [1,2,*], Ouiem Bchir [1] and Mohamed Maher Ben Ismail [1]

[1] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[2] Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
* Correspondence: saltuwayjiri@seu.edu.sa

Abstract: Automatic speaker verification (ASV) systems play a prominent role in the security field due to the usability of voice biometrics compared to alternative biometric authentication modalities. Nevertheless, ASV systems are susceptible to malicious voice spoofing attacks. In response to such threats, countermeasures have been devised to prevent breaches and ensure the safety of user data by categorizing utterances as either genuine or spoofed. In this paper, we propose a new voice spoofing countermeasure that seeks to improve the generalization of supervised learning models. This is accomplished by alleviating the problem of intraclass variance. Specifically, the proposed approach addresses the generalization challenge by splitting the classification problem into a set of local subproblems in order to lessen the supervised learning task. The system outperformed existing state-of-the-art approaches with an EER of 0.097% on the ASVspoof challenge corpora related to replaying spoofing attacks.

Keywords: speaker verification; voice spoofing; countermeasures; generalization; replay spoofing

check for updates

Citation: Altuwayjiri, S.M.; Bchir, O.; Ismail, M.M.B. Generalized Replay Spoofing Countermeasure Based on Combining Local Subclassification Models. *Appl. Sci.* **2022**, *12*, 11742. https://doi.org/10.3390/app122211742

Academic Editors: Konstantinos Rantos, Konstantinos Demertzis and George Drosatos

Received: 14 October 2022
Accepted: 15 November 2022
Published: 18 November 2022

1. Introduction

In order to protect the applications and stored data, biometric authentication is currently widely used along with other identification modalities to supervise and control system accessibility [1]. Speaker verification (SV) systems exploit speech modalities to identify the user seeking to gain access to systems or services. Specifically, human voiceprint authentication is performed by comparing the voice of the speaker to previously recorded voiceprints. The growing popularity of voice-activated smart home systems has increased the prominence of automatic speaker verification (ASV) technology as a security measure for such devices. These ASV systems also benefit other services such as phone banking and online payment processing. Nevertheless, serious security concerns constrain the potential of these systems. Indeed, spoofing attacks pose a threat to ASV systems [2]. Both the International Electrotechnical Commission (IEC) and the International Organization for Standardization (ISO) have defined such intrusions as presentation attacks (Pas) [3]. These attacks are conducted by criminals impersonating an authenticated user to attempt to gain access to private information [1], and are performed through the use of speech synthesis (SS), replay attacks, and data voice conversion (VC) techniques [1,2]. Among these techniques, replay attacks are the most common since they do not require substantial technological knowledge. Furthermore, it is difficult to detect such attacks due to the simplicity of the technique, which consists of collecting voice samples and then replaying them. To block these spoofing attacks, it is necessary to devise antispoofing countermeasures. This approach consists of using a classification system that distinguishes between genuine and spoofed utterances.

Classifying voice utterances as genuine or spoofed typically entails both performing suitable feature extraction and applying classification technique on these features. In this regard, potential audio features have been explored. In particular, Q Cepstral coefficients (CQCC) [4] and the linear frequency Cepstral coefficient (LFCC) [5] have been explored. Alternatively, for the supervised learning model, Gaussian mixture models (GMMs) [6] and deep neural networks (DNNs) [7] have been utilized in the design of spoofing countermeasures. Specifically, several deep learning architectures have been adapted to serve antispoofing purposes, such as residual neural networks (ResNets) [8] and recurrent neural networks (RNNs) [9]. Nevertheless, these models are outperformed by Gaussian mixture model (GMM)-based techniques, which were even more successful than existing industry-leading approaches [2].

Despite these advances, the trained models perform poorly when classifying new instances. This is known as the generalization problem, and it represents a significant challenge for spoofing counter-measures [2]. This problem is mainly due to the large variance of the intra-class instances [10,11]. For genuine utterances, disparities in the user voice represent the primary reason for this impediment. Indeed, the voice can be distorted by many factors such as emotional state, user health, or the authentication device, in addition to the obvious discrepancy between different users' voiceprints [12,13]. While the same variation factors apply to spoofed utterances, the latter feature additional forms of variation, such as those caused by recording devices or the algorithms deployed for manipulation purposes [2].

Given the above factors, the generalization problem can be addressed by handling the variance in both genuine and spoofed utterances. In this regard, we propose to tackle the variance problem by dividing the classification problem into several subproblems. This is performed by learning the hidden partitions of the utterances through the use of clustering techniques. Then, a countermeasure suitable for each subgroup is devised. For this purpose, various classifiers are investigated with respect to each subgroup. Furthermore, ensemble learning is exploited to seek better predictive performance.

2. Background

Both unsupervised and supervised learning paradigms are involved in the design of the proposed approach. Indeed, unsupervised learning, specifically clustering, is utilized to learn the underlying structure of the data and split it automatically into homogeneous subgroups. Alternatively, supervised learning, particularly classification, is exploited to devise a set of countermeasures suitable for each subgroup.

2.1. Clustering

Clustering uses a specific measure to group similar utterances to the same cluster and dissimilar ones to different ones. This method entails three main techniques. The first technique is hierarchical clustering, which involves establishing a hierarchal structure of the clusters by adopting either a top–down approach (known as divisive) or a bottom–up approach (known as agglomerative). The second technique is partitioning clustering (also known as centroid-based clustering). This method learns a representative instance from each cluster (e.g., the cluster centers) and assigns the instances to the closest representative. The third technique is density-based clustering [12], which assigns instances to each cluster on the basis of density. More specifically, clusters are formed of dense instances, while sparse instances are categorized as outliers.

These clustering techniques can be either crisp or fuzzy. The former involves assigning instances to only one cluster, whereas the latter uses a membership degree to assign instances to multiple clusters on the basis of their probability to belong to each cluster. This allows for fuzzy clustering to be applied to real-word problems with overlapping cluster boundaries [6]. Three fuzzy clustering processes are outlined in the next section: competitive agglomeration CA [13] algorithms, fuzzy C-means (FCM) clustering [14], and simultaneous clustering and attribute discrimination (SCAD) [15].

2.1.1. Fuzzy C-Means

By minimizing intracluster distances, fuzzy C-means (FCM) [14] conducts the fuzzy partitioning of unlabeled data. To be specific, if x_j represents a set of instances; by minimizing the objective function defined in (1) subject to (2), the cluster representatives (centers), c_i, and fuzzy memberships, (μ_{ij}) are derived. Both c_i and (μ_{ij}) are then learned alternatively through iterative learning as follows.

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij})^m \|x_j - c_i\|^2, \tag{1}$$

subject to

$$\mu_{ij} \in [0,1]\ \forall i, j;\ and\ \sum_{i=1}^{C} \mu_{ij} = 1 \tag{2}$$

In (2), d denotes the dimension of the vectors, C denotes the number of clusters, m is parameter controlling the membership fuzziness, and N is the number of utternaces, x_j and $c_i \in \mathbb{R}^d$. FCM is fast and robust with a time complexity of O(N).

2.1.2. Simultaneous Clustering and Attribute Discrimination

Feature selection and aggregation can be performed using an extension of FCM called simultaneous clustering and attribute discrimination (SCAD) [15]. For each cluster, this process learns relevant feature weights, $V = [v_{ik}]_{\substack{i\,=\,1..C,\\ k\,=\,1..d}}$ and fuzzy memberships, $U = [u_{ij}]_{\substack{i\,=\,1..C\\ j\,=\,1..N}}$ and centers, $C = [c_{ik}]_{\substack{i\,=\,1..C\\ k\,=\,1..d}}$ by minimizing the following objective function in (3):

$$J(C,\ U,\ V; X) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m \sum_{k=1}^{d} v_{ik} \left(x_{jk} - c_{ik}\right)^2 \tag{3}$$

subject to

$$0 \le u_{ij} \le 1;\ and\ \sum_{i=1}^{C} u_{ij} = 1, \tag{4}$$

and

$$v_{ik} \in [0,1]\ \forall i, k;\ and\ \sum_{k=1}^{d} v_{ik} = 1,\ \forall\ i \tag{5}$$

where N denotes the number of utterances; C denotes the cluster number; d denotes the size of the feature, with v_{ik}, c_{ik}, and $u_{iji} \in \mathbb{R}^d$. Since SCAD is based on fuzzy C-means, it is fast and robust. It has also the same time complexity as that of O(N).

2.1.3. Competitive Agglomeration

Another extension of FCM is competitive agglomeration (CA) [13]. CA handles the challenge of determining the cluster number. This technique fuses hierarchical and partitioning processes to utilize the advantages of both in order to learn the number of clusters, cluster representatives, and fuzzy memberships. CA learns the optimal cluster number by splitting the utterances into tiny groups that subsequently compete over instances in the optimization process. Consequently, empty clusters slowly vanish. CA is achieved through the optimization of the objective function in (6):

$$F = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^2 \cdot d_{ij}^2(x_j,\ \beta_i) - \alpha \sum_{i=1}^{C} \left[\sum_{j=1}^{N} u_{ij}\right]^2 \tag{6}$$

where the cluster representatives are $B = (\beta_1, \ldots, \beta_c)$, and the distance between the feature vector x_j and prototype β_i is $d_{ij}^2(x_j, \beta_i)$. u_{ij} represents the degree of belongness of utterance j to the partition i.

The cost defined in (6) contains two parts. The left term represents the FCM clustering technique as defined in (1) responsible for the fuzzy portioning, while the right term expresses the competition between instances to be enclosed in cluster competition. Similarly, CA is based on fuzzy C-means. Therefore, it is a time complexity of O(N).

2.2. Classification

Classification is a supervised learning technique where a model is built using labeled data instances in order to predict the class value for unseen instances [16]. Specifically, the model learns how to map input instances to the predefined classes. Thus, for the learned model to be effective, the set of training data should be representative and sufficiently available. A problem that requires classification can be a binary classification problem or multiclass classification problem. For binary classification, only two classes are considered, while for multi class classification, more than two classes are considered. Another way of categorizing the classification problem is as linear or nonlinear. Linear classifiers employ linear models for class prediction. Alternatively, nonlinear classifiers learn nonlinear models [6]. In the literature, various classifications algorithms have been proposed. However, there is no way to know which classification model is more suitable for a certain problem. As such, the choice of the classifier is generally empirically performed. In the following, we outline the classification approaches that are exploited in the design of the proposed approach: the Gaussian mixture model (GMM) classifier [6], support vector machine SVM) [17], and extreme gradient boosting (XGBoost) [18].

2.2.1. Gaussian Mixture Model

The Gaussian mixture model (GMM) classifier [6] learns a probabilistic model. The latter estimates an instance as a mixture of weighted Gaussians. More specifically, on the basis of the probability density functions of an input instance with respect to each class, this classifier predicts the class of each instance using Bayes' rule [19]. The mean, standard deviation, and weight of each Gaussian involved in the mixture are estimated using the expectation maximization (EM) [20] iterative approach or maximum a posteriori (MAP) approach [21].

2.2.2. Support Vector Machine

Support vector machine (SVM) [17] is a binary classifier that learns a hyperplane that separates the two considered classes. The model of the hyperplane is learned in a way that ensures a maximal margin of separation between the two classes. This version of SVM, which does not allow for any instance to reside within the margin, is called hard SVM [22]. Alternatively, in order to learn a less complex hyperplane and avoid overfitting, soft SVM [23] allows few classification errors by letting certain instances from both classes to reside within the margin. Although SVM is designed to be a binary classifier, it can be employed for multiclass problems. This application involves learning a hyperplane with respect to each class. The problem then amounts to classifying each category against all other classes, with the name versus all SVMs [24]. In case the data are not linearly separable, kernel SVM [25] is more suitable. This method applies the kernel trick to transform the data by expressing it in new space of higher dimension, allowing for a better separation of the categories.

2.2.3. Extreme Gradient Boosting

Decision tree (DT) is a classification model which consists of nested "if/else" conditions. Alternatively, a gradient boosting decision tree (GBDT) fuses a set of DT models to achieve a better DT model. The improvement of one DT model is achieved through combination with other DT models. This process involves building a series of DT models iteratively, where each new generated model accounts for and addresses the previous model's flaws. As a result, the output consists of a weighted sum of all considered DT outputs. Similarly, extreme gradient boosting (XGBoost) [18] is a GBDT. Nevertheless, it performs parallel tree boosting rather than sequential boosting like GBDT, and checks all gradient values to assess each conceivable split of the training set.

3. Related Works

Audio classification seeks learning a model that is able to predict the category of unknown audio utterance [26]. This machine learning task can benefit many practical fields such as medical applications related to diagnosing sleep bruxism [27], dementia [28], and depression [29]. Moreover, industrial applications exploited audio classification techniques for several scenarios, such as detecting machine chatter [30] and the condition of rotating machines [31] Furthermore, environmental sound recognition [32,33] has contributed to the understanding of the context of the occurring audio. In fact, it is crucial to trigger decisive actions such as evacuating a building when an alarm occurs or reaching a baby when he cries.

Recently, the ASVspoof challenge series (https://www.asvspoof.org, accessed on 10 October 2022) deployed antispoofing benchmarks [34–36]. This triggered the research on presentation attack detection [2], particularly speech synthesis (SS), voice conversion (VC), and replay attacks.

Typically, state-of-the-art approaches consist of classifying a voice utterance as genuine or spoofed. These approaches are based on conventional classification paradigms, deep learning paradigms, or a combination of supervised and unsupervised learning paradigms.

3.1. Conventional Approaches

Conventional approaches consist of two main aspects. While the first involves extracting an audio feature suitable for discriminating genuine from spoofed utterances, the second component trains a model able to categorize the extracted features. In particular, the work in [4] employed the Cepstral coefficient (CQCC) feature [4] and employed the Gaussian mixture model (GMM) [6] as a classifier. This system is considered to be a baseline approach to assess antispoofing systems [1,2]. Similarly, the countermeasure proposed in [37] extracted a combination of cochlear filter Cepstral coefficients (CFCCs) [38] and the instantaneous frequency (IF) [39], and fed them into a GMM classifier. Alternatively, the authors in [5] proposed a countermeasure based on LFCC [40] features and a GMM classifier after comparing 19 different features coupled with SVM [17] and GMM [6,41] classifiers. On the other hand, the study in [42] combined mel-frequency Cepstral coefficient (MFCC) [43], mel-frequency principal coefficient (MFPC) [44], and CosPhase principal coefficient (CosPhasePC) [45] features and conveyed them to an SVM classifier.

3.2. Deep Learning Approaches

Due to the boost achieved by deep neural networks (DNNs) in the machine learning field, particularly in classification tasks, antispoofing approaches based on a deep learning paradigm have been proposed. For this purpose, several DL models have been exploited. More specifically, the system outlined in [46] utilizes a dilated residual network (DRN) deep learning model [47] including a ResNet [47] model with an attention filtering mechanism to discard irrelevant audio segments such as background noise. The ResNet [47] deep learning model was also utilized in the system described in [48]. Here, two low-level cepstral features, MFCCs [43] and CQCCs [4], were fed into the network instead of the raw data. A similar model was deployed in the system described in [49]. However, rather than conveying MFCCs [43] as input, this model uses high-frequency Cepstral coefficients (HFCCs). Similarly, the authors in [50] employed ResNet along with SENet [51], Mean-Std ResNet [51], and Dilated ResNet [52] to analyze CQCCs and spectrogram features. The fusion in these models is performed using the greedy fusion scheme presented in [53]. As a result, the fusion of these deep learning models was found in [50] to yield a system that outperformed reported state-of-the-art approaches when using the Asvspoof 2019 Replay Benchmark.

Recurrent neural networks (RNNs) [9] have also been exploited to design countermeasures. As such, the research in [54,55] employed long short-term memory (LSTM) [56]. The research in [57,58] exploited RNN [9] along with a convolutional neural network (CNN) [59]. In these works, CNN functioned as a feature extractor, while RNN performed long dependency processing. Similarly, the study in [60] exploited a combination of CNN and RNN. Specifically, the study combined three i-vector [61] systems, namely, the light convolutional neural network (LCNN) [62] system and the CNN + RNN one. LCNN was also employed along with a small Bayesian neural network [63] in [63,64]. In [65], the softmax function was replaced with the softpus function to estimate the deep learning model prediction uncertainty. Alternatively, a light convolutional gated recurrent neural network (LC-GRNN) was used in [66], and the authors in [67] adopted a variety of LCNNs based on context gate CNN (CGCNN), which used gated linear unit (GLU) activations as a context-gate for each filter. The adopted feature for this system was Log-CQT.

Recently, variational autoencoders (VAEs) [68] have also been adopted to devise new countermeasure, such as those used in [69]. Additionally, a combination of two visual geometry group (VGG) [70] models were employed in [71].

3.3. Combination of Unsupervised and Supervised Learning

Recently, a spoofing countermeasure based on mining hidden partitions of genuine and spoofed utterances using fuzzy clustering was proposed in [72]. This countermeasure partitions each class (genuine/spoofing) into subgroups such that each subgroup shared the same characteristics and thus exhibited low variance. The classification of unknown utterances was than performed by assigning them to the closest subgroup.

Figure 1 depicts the spoofing countermeasure reported in [72]. First, audio features are extracted from all utterances. Then, the instances of each category (genuine/spoofing) are clustered using the fuzzy clustering approach. As such, the representatives of the genuine sub-categories and those of the spoofing one are learned. In particular, fuzzy clustering techniques are employed. The experimental results showed that two genuine clusters and two spoofing clusters dramatically increased the performance. It yielded a testing EER of 1.07% on the ASVspoof 2017 replay benchmark dataset.

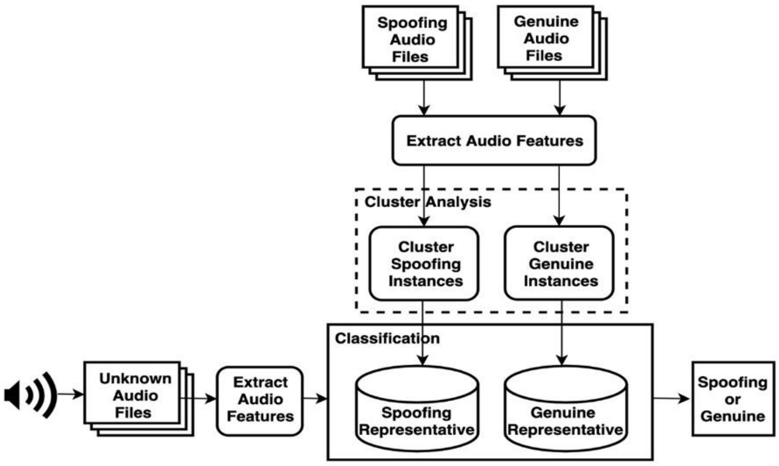

Figure 1. Architecture of the spoofing countermeasure based on mining hidden partitions of genuine and spoofed utterances [72].

An illustrative example of the countermeasure reported in [72] is depicted in Figure 2. In this example, six clusters $\{S_1, S_2, S_3, S_4, S_5, S_6\}$ are learned for the spoofing class, and four clusters $\{G_1, G_2, G_3, G_4\}$ are learned for the genuine class. Then, an unknown utterance is compared to the 10 cluster centers. Since the closest cluster is G_1, one of the genuine clusters, the unknown utterance is classified as genuine.

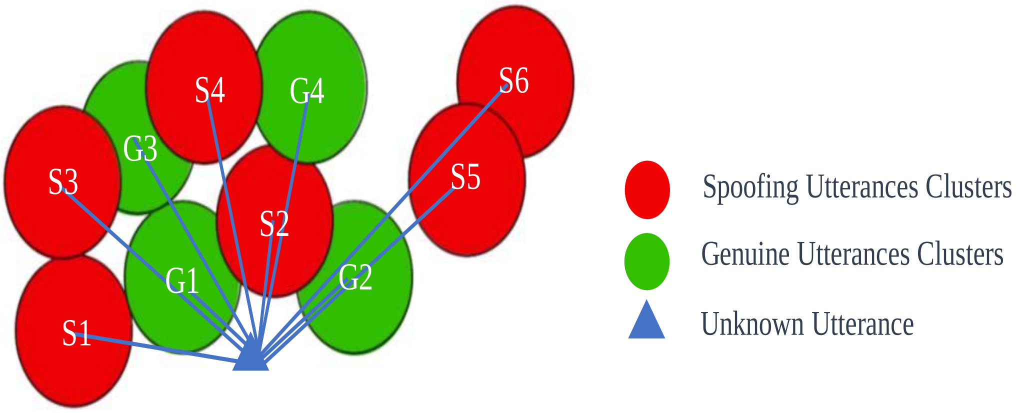

Figure 2. Illustrative example of the spoofing countermeasure based on mining hidden partitions of genuine and spoofed utterances [72].

Tables 1 and 2 report the performance of the state-of-the-art approaches for 2017 and the 2019 ASVspoof replay benchmark datasets, respectively. The study in [72] achieved the best performance, with a testing error rate of 1.07% on the ASVspoof 2017 replay benchmark dataset. For the ASVspoof 2019 replay benchmark dataset, the countermeasure proposed in [50] obtained the best performance, with a testing error rate of 0.59%.

Table 1. State-of-the-art training and testing error rates when using the ASVspoof 2017 replay benchmark dataset.

Reference	Input	Model	Training Error Rate (%)	Testing Error Rate (%)
[4]	CQCCs	GMM	10.35	24.77
[46]	Signal Logspec via FFT	ResNet	6.09	8.54
[48]	CQCC and MFCC	GMM, ResNet	2.58	13.30
[49]	Fusion of HFCC and CQCC	DNN, SVM	7.6	11.5
[54]	MFCC, Fbank	LSTM and GRU RNN	6.32	9.81
[60]	CQT and FFT	LCNN, SVM, CNN + RNN	3.95	6.73
[66]	Spectrogram features	LC-GRNN + PLDA	3.26	6.08
[69]	CQCC	C-VAE	18.1	28.1
[69]	Spectrogram	C-VAE	22.81	29.52
[73]	CQCC	LCNN	21.73	8.20
[72]	MFCC, CQCC	SCAD, KNN	0.13	**1.07**

Bold number indicates the lowest EER value.

Table 2. State-of-the-art training and testing error rates when using the ASVspoof 2019 replay benchmark dataset.

Reference	Input	Model	Training Error Rate (%)	Testing Error Rate (%)
[50]	CQCC, spectrogram	Fusion of SENet, Mean-Std ResNet, and Dilated ResNet	0.129	**0.59**
[64]	Spectrogram features	Ensemble Weights for Bayesian NN, LCNN	0.78	0.88
[66]	Spectrogram features	LC-GRNN + PLDA	0.73	2.23
[69]	CQCC	AC-VAE2	34.06	36.66
[71]	spectrogram, CQT	VGG, SincNet, LCNN	0.66	1.51

Table 2. *Cont.*

Reference	Input	Model	Training Error Rate (%)	Testing Error Rate (%)
[67]	LMS + LogCQT	LCNN	0.16	1.16
[74]	CQCC	GMM	9.87	11.04

Bold number indicates the lowest EER value.

3.4. Discussion

Neither conventional nor deep learning approaches have managed to overcome the challenge posed by the high variation of utterances. Indeed, these models suffer from generalization issues. In other words, while these countermeasures increase the prediction performance of trained utterances, they are unable to generalize utterances. Alternatively, the countermeasure proposed in [72] addressed the generalization problem by mining hidden partitions of the genuine and spoofed utterances separately. Nevertheless, while taking into account the intra-class variance by learning the underline structure of each class, this solution did not consider overlaps between genuine and spoofed categories. Indeed, this method did not learn the overall underlying structure of the data.

4. Proposed Approach: Generalized Replay Spoofing Countermeasure Based on Combining Local Sub-Classification Models

We propose an alternative approach that mines the hidden structure of the whole data. More specifically, the proposed countermeasure splits the classification problem into local subproblems. In other words, in order to avoid learning a complex classification model for the whole data, we intend to split the data into groups formed of congregated instances and build a simpler classification model from each group. These groups are heterogeneous and include spoofing and genuine utterances assigned to the same group due to their similarities. By classifying the utterances of each cluster into spoofing and genuine, a classification model is learned with respect to each cluster. This results in a set of local classification models. Using these models, the classification of an unknown instance is then achieved through an ensemble learning approach that combines the obtained local models.

The proposed spoofing countermeasure is depicted in Figure 3. First, audio features are extracted from the recorded utterances. Then, the three clustering techniques of FCM [14], SCAD [15], and CA [13] are investigated to partition the data. FCM-based clustering approaches are explored because they learn the cluster centers while also learning a fuzzy partition of the data. Alternatively, SCAD has the advantage of learning relevant feature weights and their combinations while clustering the data, whereas CA learns the number of homogeneous partitions automatically. From each cluster containing both spoofed and genuine instances, a classification model is learned. We propose to employ GMM [6] and SVM [17] as classification techniques, since these models were effective in the prediction of spoofed utterances [4,5,37,42,48,49,60]. Lastly, an ensemble learning technique is adopted to classify unknown instances by combining the decisions of the learned models. More specifically, the pairwise distances between the unknown utterance and the cluster representatives are computed. The classification model corresponding to the closest sub-group is then used for classifying this utterance.

To better illustrate the proposed spoofing countermeasure based on local classification subproblems, an example is presented in Figure 4. In this example, audio instances are clustered into six groups: $\{R_1, R_2, R_3, R_4, R_5, R_6\}$. Although the training set was labeled into spoofing and genuine instances, these labels were not used for the clustering task. In fact, the whole data were considered without consideration of the ground truth. Therefore, each obtained cluster included both spoofing and genuine instances. Then, during the training phase, a classification model was learned from each cluster. This resulted in a set of six classification models, which were used to classify the unknown utterance through ensemble learning techniques. For example, since the unknown utterance is closest to cluster R_1, the model learned for R_1 will be employed for its classification. Moreover, for each subgroup, a set of classifiers were investigated. This set contained the Gaussian nixture nodel (GMM) classifier [6], support vector machine (SVM) [17], and XGBoost [18]. To minimize learning errors and enhance the overall learning performance of each local subproblem, ensemble learning [75] was exploited to combine the considered classification results. For this purpose, the majority strategy was employed [75].

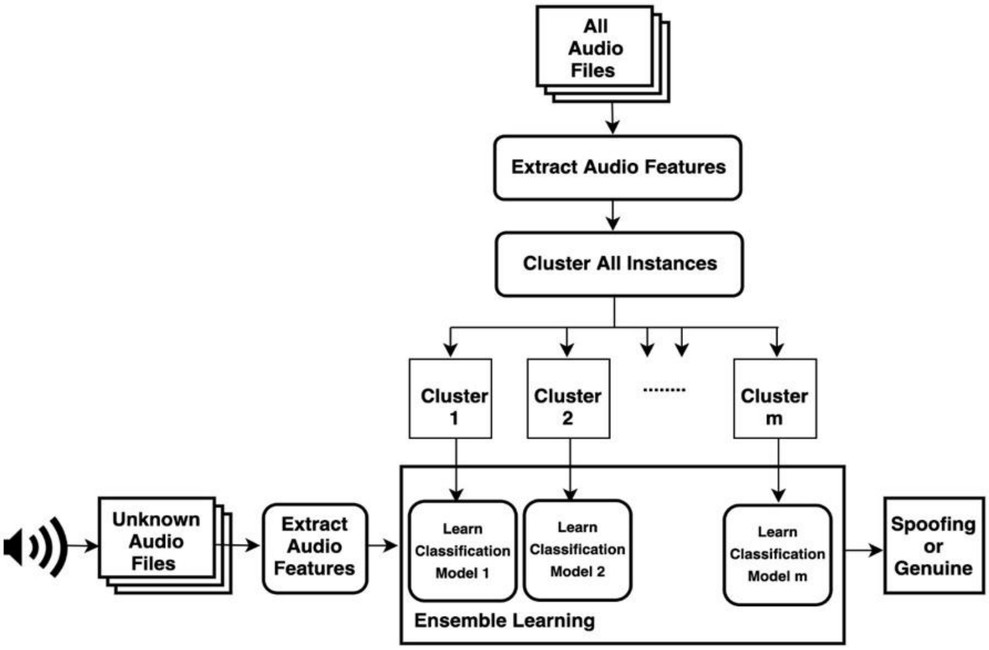

Figure 3. Architecture of the proposed spoofing countermeasure based on local classification subproblems.

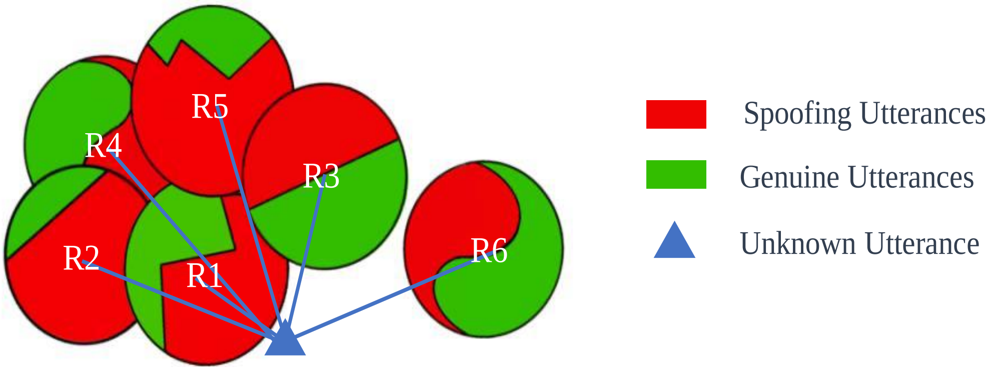

Figure 4. Illustrative example of the proposed spoofing countermeasure based on local classification subproblems.

5. Experiments

To assess the performance of the proposed approach, two replay datasets were considered: the ASVspoof 2017 version 2.0 benchmark dataset [35] and the ASVspoof 2019 benchmark dataset [36]. The audio files included in these datasets are characterized by a 16 kHz sampling rate and 16-bit resolution. As reported in Table 3, ASVspoof 2017 v2.0 was split into three subsets. The first subset was a training set that contained 3016 files, of which 1507 were genuine, and 1507 were replay spoofing files. The second subset was a development set containing 1710 files, of which 760 were genuine, and 950 were replay spoofing files. The third subset was an evaluation set containing 13,306 files, of which 1298 were genuine, and 12,008 were replay spoofing files.

Table 3. ASVspoof 2017 V.2.0 dataset.

Subset	No. of Speakers	No. of Utterances	
		Genuine	Spoofed
Training	10	1507	1507
Development	8	760	950
Evaluation	24	1298	12,008

As shown in Table 4, the ASVspoof 2019 replay spoofing dataset comprised a training set with 48,600 spoofed utterances and 5400 genuine utterances, a development set with 24,300 spoofed utterances and 5400 genuine utterances, and an evaluation set containing various randomly chosen acoustic and playback configurations [36].

Table 4. ASVspoof 2019 replay spoofing dataset.

Subset	No. of Speakers	No. of Utterances	
		Genuine	Spoofed
Training	20	5400	48,600
Development	20	5400	24,300
Evaluation	-	137,457	

From the audio files, three audio features were extracted: mel-frequency Cepstral coefficients (MFCCs) [43], the constant Q Cepstral coefficients (CQCCs) [4], and the linear frequency Cepstral coefficient (LFCC) [5]. The equal error rate (EER) [76] is considered as the performance measure. EER represents the operating point at which the false acceptance rate (FAR) and false rejection rate (FRR) are equal [76].

5.1. Experiment 1: Number of Clusters and Audio Feature Investigation

In this experiment, the FCM [14] clustering approach was employed to mine the hidden structure of the data. This approach partitions the whole ASVspoof 2017 benchmark dataset into homogeneous local subgroups. Each subgroup contained genuine and replay spoofed utterances, the latter of which constituted a local classification subproblem. Two classifiers, SVM [17] with linear kernel and GMM [6] with two mixture components, were utilized to solve these subproblems. Furthermore, to explore the structure of the data, different numbers of clusters were considered. These numbers were tuned between 2 and 15. Moreover, the data were clustered using CQCC, MFCC, and LFCC features. Each feature was considered independently and concatenated together. Figures 5 and 6 depict the EER obtained with MFCC, QCC, LFCC, and concatenation together with respect to the cluster number when considering SVM and GMM classifiers, respectively. The results indicate that performance varied with respect to the number of clusters, the type of audio features, and the classifier.

For SVM-based systems, CQCC features generally performed better than the other considered features, especially when the cluster number was less than 9. However, the best performance was achieved with two clusters. Alternatively, for GMM-based approaches, the best performance was attained with four clusters. Nonetheless, CQCC remained the best performing feature type. Table 5 reports the best performance achieved by each combination of feature/classifier for the optimal number of clusters. The system that used CQCC features with two clusters had the smallest EER (1.61%) and thus outperformed the other combinations. The second best was the system employing CQCC features and GMM, with an ERR of 4.23%.

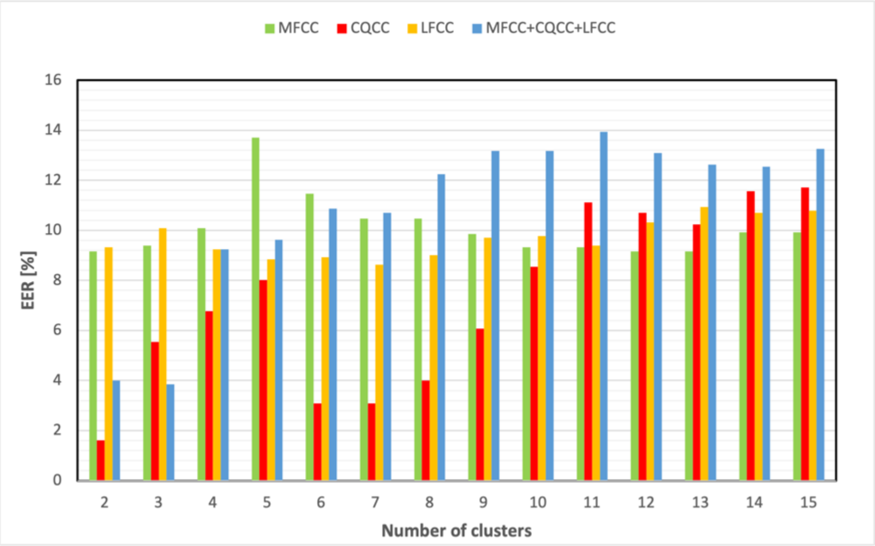

Figure 5. EER obtained with MFCC, QCC, LFCC, and concatenation together with respect to the number of clusters when using SVM on the ASVspoof 2017 version 2.0 benchmark dataset.

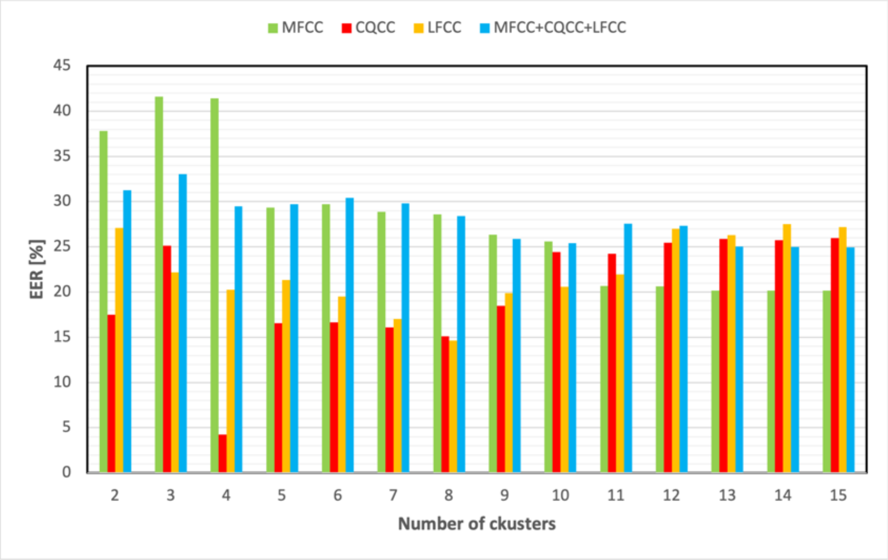

Figure 6. EER obtained by MFCC, QCC, LFCC, and concatenation together with respect to the number of clusters when using GMM on the ASVspoof 2017 version 2.0 benchmark dataset.

Table 5. Best performance achieved by each combination of feature/classifier for the optimal number of partitions on the ASVspoof 2017 benchmark dataset.

Model	No of Clusters	EER%
SVM with MFCC	2	9.16
SVM with CQCC	2	**1.61**
SVM with LFCC	7	8.62
SVM with ALL	3	3.85
GMM with MFCC	13	20.18
GMM with CQCC	4	4.23
GMM with LFCC	8	14.63

Table 5. *Cont.*

Model	No of Clusters	EER%
GMM with ALL	15	24.96

Bold number indicates the lowest EER value.

5.2. Experiment 2: Self-Learning the Number of Clusters

In this experiment, the hidden partition was discovered automatically using the competitive agglomeration (CA) [13] clustering approach to simultaneously partition the training utterances and estimate the cluster number. The cluster number was first set to 100. The ASVspoof 2017 and the ASVspoof 2019 benchmark datasets were considered in this experiment. Table 6 reports the EER obtained when employing SVM as a classifier along with the cluster number learned for each feature. As shown in Table 6, CQCC achieved the lowest EER of 1.42% and 1.63% on the ASVspoof 2017 and ASVspoof 2019 dataset, respectively, with an optimal number of clusters equal to 2. Starting from a large number of 100, CA achieved similar results to those obtained in the first experiment by tuning the number of clusters. Alternatively, Table 7 reports the obtained EERs when using GMM classifier along with the cluster number learned for each feature. The results confirm the superiority of the CQCC features, which achieved an EER of 1.38% and 1.46% on the ASVspoof 2017 and ASVspoof 2019 dataset, respectively, while learning an optimal number of clusters equal to four. This result is consistent with the results obtained by exploring the cluster number in experiment 1. This suggests that CA can discover the hidden partitions of the data while self-learning the optimal number of clusters.

Table 6. Achieved EER when employing SVM as a classifier along with the estimated cluster numbers for each feature on the ASVspoof 2017 and ASVspoof 2019 benchmark datasets.

		MFCC	CQCC	LFCC	CQCC + MFCC + LFCC
ASVspoof 2017	EER	8.24	**1.42**	7.29	3
	No. of clusters	3	2	5	2
ASVspoof 2019	EER	9.79	**1.63**	6.88	2.34
	No. of clusters	4	2	2	2

Bold number indicates the lowest EER value.

Table 7. Achieved EER when using the GMM classifier and the learned number of clusters with respect to the considered features on the ASVspoof 2017 and ASVspoof 2019 benchmark datasets.

		MFCC	CQCC	LFCC	CQCC + MFCC + LFCC
ASVspoof 2017	EER	14.4	**1.38**	9.25	9.47
	No. of clusters	3	4	4	3
ASVspoof 2019	EER	12.23	**1.46**	7	11.78
	No. of clusters	3	2	5	5

Bold number indicates the lowest EER value.

5.3. Experiment 3: Feature Relevance Weight Learning

In this experiment, simultaneous clustering and attribute discrimination (SCAD) [15] was used to mine the hidden structure of the data, and learnt the relevant feature weights of CQCC, MFCC, and LFCC. First, the cluster number was set to 2 for SVM-based systems and 4 for GMM-based systems, in accordance with the obtained results in Experiments 1 and 2. Table 8 reports the learned feature relevance weights. As shown in Table 8, the largest weight was assigned CQCC. This result is consistent with Experiment 1 findings, which proved that CQCC is more suitable.

Table 8. Feature weights estimated for each cluster on the ASVspoof 2017 version 2.0 benchmark dataset.

	CQCC	**MFCC**	**LFCC**
SCAD with SVM			
Cluster 1	0.999963	0.000003	0.000034
Cluster 2	0.999962	0.000003	0.000035
SCAD with GMM			
Cluster 1	0.999963	0.0000032	0.000033
Cluster 2	0.999962	0.0000033	0.000034
Cluster 3	0.999962	0.0000033	0.000034
Cluster 4	0.999963	0.0000032	0.000034

Next, we discarded MFCC and LFCC, and applied SCAD to the entries of CQCC to learn the relevance of each entry. The considered cluster numbers were between 2 and 16. Figure 7 shows the achieved EER for each cluster number when employing SCAD and SVM [17] on the ASVspoof 2017 version 2.0 benchmark dataset. The lowest EER, equal to 0.154, was achieved for 2 clusters. When using the GMM classifier, the lowest EER was equal to 0.302 with four clusters, as shown in Figure 8. This suggests that employing SCAD on the CQCC gave better performance because this approach handled the large dimension of CQCC feature by computing the weighted sum of the feature entrees.

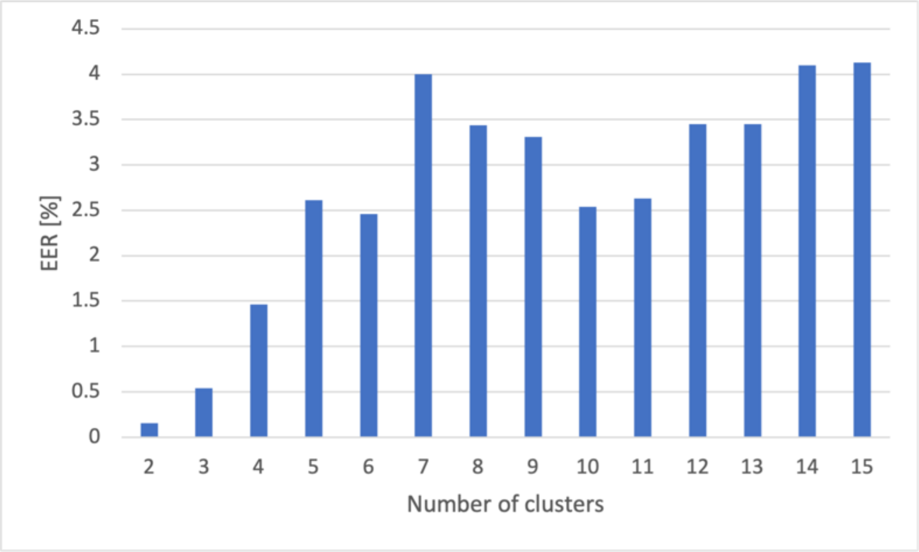

Figure 7. The achieved EER for each cluster number when using SCAD and SVM [17] on the ASVspoof 2017 version 2.0 benchmark dataset.

5.4. Ensemble Learning

On the basis of the findings of previous experiments, we only considered CQCC features in this experiment. Next, we applied the CA [13] algorithm to estimate the optimal cluster number. The learned fuzzy memberships were next used as the initial values for the SCAD [15] clustering algorithm. Three classifiers were first considered separately: SVM [17], GMM [6], and XGboost [18]. Next, the results of these classifiers were combined using the majority vote ensemble learning strategy. Table 9 depicts the achieved ERR of the considered systems. As shown in Table 9, the proposed approach based on SVM [17] outperformed those based on XGboost [18] and GMM [6] with an ERR equal to 0.154%. Furthermore, the ensemble majority voting strategy further improved performance by achieving an ERR equal to 0.097%.

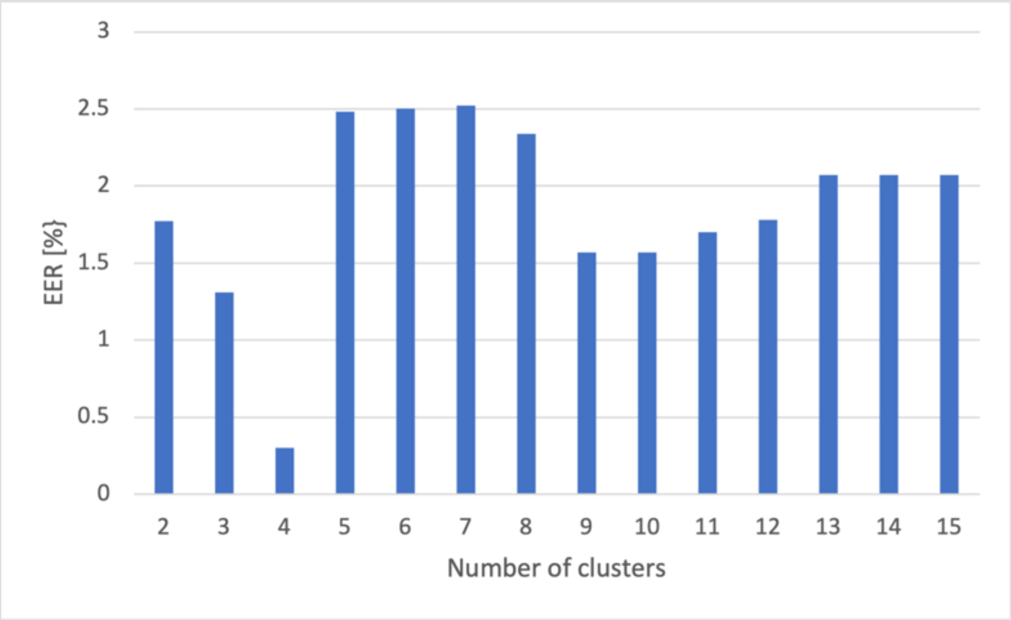

Figure 8. EER with respect to the number of clusters when using SCAD with GMM [6] on CQCC on the ASVspoof 2017 version 2.0 benchmark dataset.

Table 9. EER for SCAD with different classifiers on the ASVspoof 2017 version 2.0 Benchmark Dataset.

Model	No of Clusters	EER (%)
SCAD + SVM	2	0.154
SCAD + GMM	4	0.302
SCAD + XGboost	3	1.07
Majority vote on SVM, GMM, and XGboost	2, 4, 3	**0.097**

Bold number indicates the lowest EER value.

5.5. Experiment 4: Performance Comparison with Related Spoofing Detection Approaches

According to the previous experiments' findings, the SCAD clustering algorithm with CQCC achieved the best performance with respect to the three considered classifiers. As such, in this experiment, we considered four versions of the proposed approach using CQCC for feature extraction and SCAD for mining the structure of the data. These approaches use SVM, GMM, XGBoost, and their combination. These methods are referred to as the local-SVM-based approach, local-GMM-based approach, local-XGBoost-based approach, and local-ensemble-learning-based approach. We also compared the performance of the proposed approaches to three state-of-the-art approaches. The first was the approach reported in [4], which consisted of extracting the CQCC feature and conveying it to a GMM-based classifier. The second approach, which is the most recent, was reported in [68], and it uses SCAD to cluster genuine utterances into G clusters, with the spoofed utterances placed into two S clusters. As such, this approach assigns the unknown instance as the class of the closest cluster (refer to Section 3.3). The third baseline approach, published in [46], was the best performing method for the ASVspoof replay 2019 dataset. This approach uses CQCCS and spectrogram features and conveys them to the SENet [47], Mean-Std ResNet [47], and Dilated ResNet [48] deep-learning models. Then, the greedy fusion scheme described in [49] was employed to explore the best system combination.

For this purpose, we considered the two available replay datasets: ASVspoof 2017 v2 [72] and ASVspoof 2019 [73]. The same datasets with the same training and testing sets were employed for all considered approaches. To evaluate the generalization capabilities of the proposed approach, both the training and testing ERR were compared, as reported in Table 10, where the proposed approach based on ensemble learning outperformed all other considered systems with respect to the two datasets. Nonetheless, even without considering ensemble learning, the three other approaches achieved smaller EERs than the state-of-the-art ones, except for the local-GMM-based approach, which offered the same performance as baseline approach 3 on the ASVspoof replay 2019 dataset. This result was achieved by dividing the classification problem into sublocal problems to address the utterance high variance problem and was confirmed by the training and testing ERR results. The difference between the training and testing ERR was reduced. This result shows that the generalization problem was addressed.

Table 10. Performance comparison with state-of-the-art approaches.

Model	ASVspoof 2019			ASVspoof 2017 v2		
	No. of Clusters	Training ERR (%)	Testing ERR (%)	No. of Clusters	Training ERR (%)	Testing ERR (%)
Baseline System 1 [4]	-	9.87	11.04	-	10.35	24.77
Baseline System 2 [68]	15 for genuine and 15 for spoofing	2.30	3.31	2 for genuine and 2 for spoofing	0.13	1.07
Baseline system 3 [46]	-	0.129	**0.59**	-	-	-
Local-SVM-based approach	2	0.075	0.149	2	0.067	0.154
Local-GMM-based approach	2	0.189	0.59	4	0.151	0.302
Local-XGBoost-based approach	3	0.74	1.36	3	0.435	1.07
Local-Ensemble Learning-based approach	2, 2, 3	0.06	**0.119**	2, 4, 3	0.0385	**0.097**

Bold number indicates the lowest EER value.

6. Conclusions and Future Works

Spoofing detection approaches is crucial to protect the user data against voice spoofing attacks while using ASV. These spoofing detection approaches amount to a classification problem where audio utterances are categorized into genuine or spoofed classes. However, this task remains challenging due to the high variance of the utterances. This factor affects the model's generalization for unseen utterances.

In this paper, we devised a new replay countermeasure to address the high variance of these utterances. This countermeasure was performed by dividing the challenging classification problem into a set of local subproblems by mining the hidden structure of the data. Then, ensemble learning was used to combine these submodels. Various features, clustering techniques, classifiers, and their combinations were investigated. The experiments showed that CA clustering can automatically learn the number of homogeneous partitions of the data. Moreover, the experimental results showed that CQCC audio features along with the SVM classifier and SCAD clustering technique are the most suitable techniques to build the proposed approach. As a result, the latter method outperformed state-of-the-art approaches. Furthermore, when combining the results of the three classifiers (SVM, GMM, and XGBoost), the proposed approach achieved even better results.

In future work, other audio features, classifiers, clustering techniques, and ensemble learning strategies could be investigated. Moreover, the performance of the proposed approaches on other types of voice spoofing could be explored.

Author Contributions: Conceptualization, S.M.A., O.B. and M.M.B.I.; methodology, S.M.A., O.B. and M.M.B.I.; software, S.M.A.; validation, S.M.A., O.B. and M.M.B.I.; formal analysis, S.M.A., O.B.; investigation, S.M.A. and O.B.; resources, S.M.A.; data curation, S.M.A.; writing—original draft preparation, S.M.A.; writing—review and editing, O.B. and M.M.B.I.; visualization, S.M.A.; supervision, O.B. and M.M.B.I.; project administration, S.M.A., O.B. and M.M.B.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ASVspoof 2017 version 2.0 and ASVspoof 2019 benchmark datasets can be found at https://www.asvspoof.org/database (accessed on 1 September 2021).

Acknowledgments: This work was supported by the Research Center of the College of Computer and Information Sciences at King Saud University. The authors are grateful for this support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kamble, M.R.; Sailor, H.B.; Patil, H.A.; Li, H. Advances in anti-spoofing: From the perspective of ASVspoof challenges. *APSIPA Trans. Signal Inf. Process.* **2020**, *9*, e2 [CrossRef]
2. Sahidullah, M.; Delgado, H.; Todisco, M.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Lee, K.A. Introduction to Voice Presentation Attack Detection and Recent Advances. In *Advances in Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019.
3. Editor, P. *Draft International Standard Iso/Iec Dis 30107-3 Information Technology—Biometric Presentation Attack Detection*; International Organization for Standardization: Geneva, Switzerland, 2017.
4. Todisco, M.; Delgado, H.; Evans, N. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. *Proc. Odyssey* **2016**, *2016*, 283–290.
5. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A Comparison of Features for Synthetic Speech Detection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015.
6. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite Mixture Models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
7. Neapolitan, R.E. Neural Networks and Deep Learning. In *Artificial Intelligence: With an Introduction to Machine Learning*; CRC Press: Boca Raton, FL, USA, 2018; pp. 389–411. [CrossRef]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
9. Medsker, L.R.; Jain, L.C. *Recurrent Neural Networks Design and Applications*; CRC Press: Boca Raton, FL, USA, 2001.
10. Ajili, M. Moez Ajili To cite this version: HAL Id: Tel-01774394 Reliability of voice comparison for forensic applications. *Avignon* **2018**. [CrossRef]
11. Kwon, Y.; Chung, S.W.; Kang, H.G. Intra-class variation reduction of speaker representation in disentanglement framework. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Shanghai, China, 25–29 October 2020.
12. Campello, R.J.G.B.; Kröger, P.; Sander, J.; Zimek, A. Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, 1343. [CrossRef]
13. Frigui, H.; Krishnapuram, R. A robust competitive clustering algorithm with applications in computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 450–465. [CrossRef]
14. Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
15. Frigui, H.; Nasraoui, O. Simultaneous clustering and attribute discrimination. In Proceedings of the IEEE International Conference on Fuzzy Systems, San Antonio, TX, USA, 7–10 May 2000.
16. Zhou, Z.-H. *Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2021.
17. Boser, B.; Guyon, I.; Vapnik, V. A Training Algorithm for Optimal Margin Classifier. *Proc. Fifth Annu. ACM Work. Comput. Learn. Theory* **1996**, *5*, 144–152. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
19. Zhang, Y. The Development of Bayesian Theory and Its Applications in Business and Bioinformatics. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *128*, 12120. [CrossRef]
20. Setoodeh, P.; Habibi, S.; Haykin, S. Expectation Maximization. In *Nonlinear Filters: Theory and Applications*; John Wiley & Sons.: Hoboken, NJ, USA, 2022; pp. 185–201, ISBN 9781118835814.
21. Chehrehsa, S.; Moir, T.J. Speech enhancement using Maximum A-Posteriori and Gaussian Mixture Models for speech and noise Periodogram estimation. *Comput. Speech Lang.* **2016**, *36*, 58–71. [CrossRef]
22. Statnikov, A.; Hardin, D.; Guyon, I.; Aliferis, C. A Gentle Introduction to Support Vector Machines in Biomedicine. In *Case Studies And Benchmarks*; World Scientific Publishing Company: Singapore, 2021.
23. Funaya, H.; Ikeda, K. A statistical analysis of soft-margin support vector machines for non-separable problems. In Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, 10–15 June 2012; pp. 1–7.
24. Jay, M.; Venkataramani, B. Design of a real time automatic speech recognition system using Modified One Against All SVM classifier. *Microprocess. Microsystems—Embed. Hardw. Des.* **2011**, *35*, 568–578. [CrossRef]
25. Schölkopf, B.; Smola, A.J. *Learning with Kernels*; MIT Press: Cambridge, UK, 2000.
26. Rong, F. Audio Classification Method Based on Machine Learning. In Proceedings of the 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 17–18 December 2016; pp. 81–84.

27. Peruzzi, G.; Galli, A.; Pozzebon, A. A Novel Methodology to Remotely and Early Diagnose Sleep Bruxism by Leveraging on Audio Signals and Embedded Machine Learning. In Proceedings of the 2022 IEEE International Symposium on Measurements & Networking (M&N), Padua, Italy, 18–20 July 2022; pp. 1–6.
28. Shimoda, A.; Li, Y.; Hayashi, H.; Kondo, N. Dementia risks identified by vocal features via telephone conversations: A novel machine learning prediction model. *PLoS ONE* **2021**, *16*, 0253988. [CrossRef] [PubMed]
29. Smith, M.; Dietrich, B.J.; Bai, E.; Bockholt, H.J. Vocal pattern detection of depression among older adults. *Int. J. Ment. Health Nurs.* **2020**, *29*, 440–449. [CrossRef] [PubMed]
30. Kvinevskiy, I.; Bedi, S.; Mann, S. Detecting machine chatter using audio data and machine learning. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 3707–3716. [CrossRef]
31. Radonjic, M.; Vujnovic, S.; Krstić, A.; Zecevic, Z. IoT System for Detecting the Condition of Rotating Machines Based on Acoustic Signals. *Appl. Sci.* **2022**, *12*, 4385. [CrossRef]
32. Nanni, L.; Maguolo, G.; Brahnam, S.; Paci, M. An Ensemble of Convolutional Neural Networks for Audio Classification. *Appl. Sci.* **2021**, *11*, 5796. [CrossRef]
33. Xu, Y.; Afshar, S.; Wang, R.; Cohen, G.; Singh Thakur, C.; Hamilton, T.J.; van Schaik, A. A Biologically Inspired Sound Localisation System Using a Silicon Cochlea Pair. *Appl. Sci.* **2021**, *11*, 1519. [CrossRef]
34. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015.
35. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August 2017.
36. ASVspoof Consortium Asvspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. Available online: http://www.asvspoof.org/asvspoof2019/%0Dasvspoof2019evaluationplan.pdf (accessed on 10 August 2022).
37. Patel, T.B.; Patil, H.A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2015.
38. Li, Q. An auditory-based transfrom for audio signal processing. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 18–21 October 2009.
39. Quatieri, T.F. *Discrete-Time Speech Signal Processing: Principles and Practice*; Pearson Education Taiwan: Taipei, China, 2005; ISBN 9861541322 9789861541327.
40. Zhou, X.; Garcia-Romero, D.; Duraiswami, R.; Espy-Wilson, C.; Shamma, S. Linear Versus Mel Frequency Cepstral Coefficients for Speaker Recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*; IEEE: New York, NY, USA, 2011. [CrossRef]
41. Bishop, C.M. *Machine Learning and Pattern Recoginiton*; Springer: Berlin/Heidelberg, Germany, 2006; ISBN 9780387310732.
42. Novoselov, S.; Kozlov, A.; Lavrentyeva, G.; Simonchik, K.; Shchemelinin, V. *STC Anti-Spoofing Systems for the ASVsPoof 2015 Challenge*; IEEE: New York, NY, USA, 2016.
43. Davis, S.B.; Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoust.* **1980**, *28*, 357–366. [CrossRef]
44. Ding, P.; Zhang, L. Speaker recognition using principal component analysis. *Proc. ICONIP2001* **2001**.
45. Wu, Z.; Chng, E.; Li, H. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. In Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
46. Lai, C.-I.; Abad, A.; Richmond, K.; Yamagishi, J.; Dehak, N.; King, S. *Attentive Filtering Networks for Audio Replay Attack Detection*; IEEE: New York, NY, USA, 2018; pp. 1–5.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Hong Kong, China, 16–18 March 2016.
48. Chen, Z.; Xie, Z.; Zhang, W.; Xu, X. ResNet and model fusion for automatic spoofing detection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2017.
49. Nagarsheth, P.; Khoury, E.; Patil, K.; Garland, M. Replay attack detection using DNN for channel discrimination. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2017.
50. Lai, J.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks. *arxiv* **2019**, arXiv:1904.01120.
51. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks 2017. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
52. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
53. Villalba, J.; Chen, N.; Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Borgstrom, J.; Richardson, F.; Shon, S.; Grondin, F.; et al. *The JHU-MIT System Description for NIST SRE18*; Johns Hopkins University: Baltimore, MD, USA, 2018.

54. Chen, Z.; Zhang, W.; Xie, Z.; Xu, X.; Chen, D. Recurrent neural networks for automatic replay spoofing attack detection. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, Alberta, AB, Canada, 15–20 April 2018.
55. Scardapane, S.; Stoffl, L.; Rohrbein, F.; Uncini, A. On the use of deep recurrent neural networks for detecting audio spoofing attacks. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, Alaska, 14–19 May 2017.
56. Ushiku, Y. Long Short-Term Memory. *Comput. Vis.* **2021**, 768–773.
57. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A.; Gomez, A.M. A deep identity representation for noise robust spoofing detection. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Hyderabad, India, 2–6 September 2018.
58. Zhang, C.; Yu, C.; Hansen, J.H.L. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 684–694. [CrossRef]
59. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning—Whole Book*; MIT Press: Cambridge, MA, USA, 2016. [CrossRef]
60. Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio replay attack detection with deep learning frameworks. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September 2017.
61. Dehak, N.; Dehak, R.; Kenny, P.; Brummer, N.; Ouellet, P.; Dumouchel, P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brighton, UK, 6–10 September 2009.
62. Wu, X.; He, R.; Sun, Z.; Tan, T. A light CNN for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2884–2896. [CrossRef]
63. Gal, Y.; Ghahramani, Z. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ICLR Work.* **2016**, *1506*, 02158.
64. Białobrzeski, R.; Kośmider, M.; Matuszewski, M.; Plata, M.; Rakowski, A. Robust Bayesian and Light Neural Networks for Voice Spoofing Detection. In Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 1028–1032.
65. Sensoy, M.; Kandemir, M.; Kaplan, L. Evidential Deep Learning to Quantify Classification Uncertainty. *Adv. Neural Inf. Process. Syst.* **2018**, arXiv:abs/1806.01768*31*, 3179–3189.
66. Gomez-Alanis, A.; Peinado, A.; Gonzalez Lopez, J.; Gomez, A. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection. *Proc. Interspeech* **2019**, *2019*, 1068–1072.
67. Yang, Y.; Wang, H.; Dinkel, H.; Chen, Z.; Wang, S.; Qian, Y.; Yu, K. The SJTU Robust Anti-Spoofing System for the ASVspoof 2019 Challenge. *Interspeech* **2019**, 1038–1042. [CrossRef]
68. Lopez, R.; Boyeau, P.; Yosef, N.; Jordan, M.; Regier, J. Decision-Making with Auto-Encoding Variational Bayes. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 5081–5092.
69. Chettri, B.; Kinnunen, T.; Benetos, E. Deep Generative Variational Autoencoding for Replay Spoof Detection in Automatic Speaker Verification. *Comput. Speech Lang.* **2020**, *63*, 101092. [CrossRef]
70. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
71. Zeinali, H.; Stafylakis, T.; Athanasopoulou, G.; Rohdin, J.; Gkinis, I.; Burget, L.; Černocký, J. Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge. *arxiv* **2019**, arXiv:1907.12908.
72. Altuwayjiri, S.M.; Bchir, O.; Ismail, M.M. Ben Mining Hidden Partitions of Voice Utterances using Fuzzy Clustering for Generalized Voice Spoofing Countermeasures. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 841–849. [CrossRef]
73. Süslü, Ç.; Eren, E.; Demiroğlu, C. Uncertainty assessment for detection of spoofing attacks to speaker verification systems using a Bayesian approach. *Speech Commun.* **2022**, *137*, 44–51. [CrossRef]
74. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection 2019. *arxiv* **2019**, arXiv:1904.05441.
75. Zhou, Z.-H. Ensemble Learning. In *Machine Learning*; Springer: Singapore, 2021; pp. 181–210, ISBN 978-981-15-1967-3.
76. Schuckers, M.E. Receiver Operating Characteristic Curve and Equal Error Rate. In *Computational Methods in Biometric Authentication: Statistical Methods for Performance Evaluation*; Springer: London, UK, 2010; pp. 155–204, ISBN 978-1-84996-202-5.