Can We Trust Your Voice? Exploring Vulnerabilities in Voice Authentication

Ke Li, Cameron Baird, Dan Lin
Department of Computer Science, Vanderbilt University, Nashville, TN 37211 USA
{ke.li.1, cameron.j.baird, dan.lin}@Vanderbilt.Edu

***Abstract*—As voice authentication technology becomes more prevalent, its security flaws and vulnerabilities are garnering increasing scrutiny. State-of-the-art deep neural network (DNN) systems for voice authentication can achieve an accuracy of over 95%. However, DNN-based models are known to be vulnerable to attacks such as adversarial examples and data poisoning. An adversary may also take advantage of the limited generalization of current DNN-based models to circumvent the system, only requiring authentic voices to impersonate others. In this paper, we leverage a data poisoning attack and two voice authentication models to investigate the vulnerability and corresponding impacts on individual user and system security. We introduce a new toolkit, the Voice Authentication Poisoning Impact Evaluator (VAPIE), incorporating conventional machine learning and VGG-based models. VAPIE is designed to predict the potential impacts of various data poisoning scenarios launched by different attackers and to evaluate overall system security, achieving an accuracy rate of over 70%. This facilitates a deeper understanding and mitigation of the risks associated with voice authentication technologies.**

***Index Terms*—Voice Authentication, Deep Neural Network, Data Poisoning Attack, SVM, Random Forest, VGG19**

I. Introduction

Voice authentication (VA), also known as speaker verification, is a biometric technique for verifying an individual's claimed identity using the unique features of their voice. VA can be more convenient than passwords and PIN codes and it has gained popularity in various applications [1], [2]. It is envisioned that in the near future users will be able to use VA to log into personal and shared devices. Currently, DNN-based models for VA can authenticate users with accuracy as high as 95% [3]–[5]. However, an overlooked vulnerability exists in current VA systems: the underlying neural network may recognize different people as the same person. For example, consider Figure 1. If two users $u1$ and $u2$ have similar voices, $u2$ may gain access to the account of $u1$ [6], [7]. This vulnerability in VA systems may be further magnified after data poisoning attacks. The $Adversary\ A$ in Figure 1 represents this concern. The data poisoning attack intentionally corrupts the training data of a machine learning algorithm to compromise the integrity and effectiveness of the model [7], [8].

In this paper, we explore this vulnerability in VA systems and propose a novel method to predict the impact of such vulnerability under data poisoning attacks launched by different attackers. Specifically, our research is guided by several intriguing questions: Can users access others' accounts using their voice? How many can they access? Are particular users inherently more effective at impersonating multiple accounts? What are the common features of such users? Furthermore, we investigate whether a data poisoning attack enhances an attacker's ability to access more accounts and affects overall voice authentication success rates.

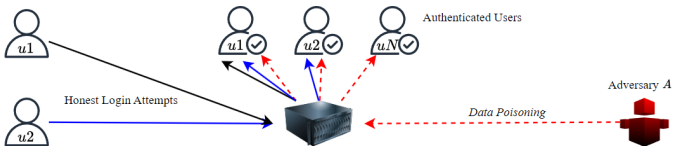


Fig. 1: Vulnerability & Data Poisoning Attack.

To answer the above questions, we have conducted several experiments using the state-of-the-art voice authentication models, DeepSpeaker [4] and ResNet [5], which yield above 95% recognition accuracy. We first examine the vulnerabilities of voice authentication models by calculating the number of potential imposters and their capabilities. Then, we launch data poisoning attacks on DeepSpeaker and evaluate the effect of the attack. Finally, we propose a machine learning based toolkit (including conventional machine learning models and VGG-based models), called Voice Authentication Poisoning Impact Evaluator (VAPIE), to predict the impact of possible data poisoning attacks from various attackers. The VAPIE model is a valuable tool for assessing the security of a voice authentication system prior to deployment.

The rest of the paper is organized as follows. Section II reviews related work. Section III presents our proposed method. Section IV reports the experimental results. Finally, Section V concludes the paper.

II. Related Work

A. Overall Voice Authentication (VA) Process

Consider a login system, such as online banking, that employs a VA system. The process of authenticating users with some system is generally called Automatic Speaker Verification (ASV) [1], [9]. Note the distinction between speaker *verification* and speaker *recognition* [10], where each speech sample is predicted to be from one of N speakers. Modern ASV systems use deep neural networks (DNNs) to extract unique features/embeddings for each new speaker that registers into the system. Though speakers can vary their content and language, characteristics like rhythm and cadence

are somewhat consistent. DNNs trained on speech data learn to map speech samples to high-dimensional features/embeddings that act as digital ID for that speaker.

The VA process begins with the registration phase. When a new user signs up for a VA account, they will be prompted to upload speech samples to the system. The neural network embeddings corresponding to these samples are stored as reference for later login attempts. After registration, the user provides new speech samples as login attempts to their account. To check if the user should be verified, the features corresponding to the new sample are compared with the reference speech using distance functions like cosine similarity [10].

To compute feature embeddings from speech samples, DNNs are the state-of-the-art solution. Early approaches used statistical features [10] or parametric approaches [11], [12]. However, modern DNNs outperform classical machine learning techniques when applied to VA [3]–[5], [13]. For this reason, we chose a DNN-based model for experiments.

B. Data Poisoning Attacks

In a data poisoning attack [14], an adversary tampers with training data to intentionally mislead a machine learning model. A recent industry survey [15] found that data poisoning may pose more of a security threat than other types of machine learning attacks such as model inversion or adversarial examples. There are several types of data poisoning. A *backdoor attack* [16] is characterized by a specific trigger pattern that is embedded into the training data. When that trigger, or backdoor, is present during inference, the model will produce a predetermined output. Backdoor attacks can be subtle as trigger patterns are often small perturbations to the input that are not noticeable by humans [17]. Another type of data poisoning attack is *label flipping* [18], where the labels of some training examples are switched to incorrect ones.

Although some works about data poisoning can be borrowed from the image domain [19], [20], there are generally less works for audio. In addition, methods used in the image domain [21], [22] will not necessarily transfer to speech [23]. However, there are several relevant works about data poisoning for voice authentication models. Guo et al. [6] present a black-box backdoor attack on real-world VA systems, showing that the backdoor can be used as a "master key" to log in to an arbitrary user account. Additionally, Zhai et al. [24] proposed a backdoor attack based on clustering of speakers in the training dataset, launching further successful attacks against open-source ASV systems. Overall, VA still has many security flaws. While some preliminary defenses exist [20], [25], [26], successful attacks on real-world systems pose critical concerns for the future of speaker verification.

III. Our Proposed Research

In this section, we first present our research findings with respect to the following research questions, and then introduce our proposed Voice Authentication Poisoning Impact Evaluator (VAPIE).

- RQ1: How many benign users can access other users' accounts via voice authentication?
- RQ2: Are some users' voices more capable of accessing more other users' accounts via voice authentication?
- RQ3: Will a data poisoning attack significantly drop the overall voice authentication success rate and hence be noticed by the voice authentication system?
- RQ4: Will an attacker gain more imposture access after a data poisoning attack?
- RQ5: Will some benign users also gain more imposture access as a side effect of a data poisoning attack by an attacker?

We address the first two research questions by evaluating the advanced voice authentication model, DeepSpeaker [4] and ResNet [5].

A. Identifying Vulnerabilities in Voice Authentication Models

We utilized the LibriSpeech dataset [27] to develop and assess our voice authentication models, selecting three partitions: "train-clean-100", "train-clean-360" and "train-other-500". We merged "train-clean-100" and "train-clean-360" for training and used "train-other-500" for evaluation. Each training partition was trimmed to 10 audio files per user, and the evaluation partition to 20, with the first 10 files serving as authentication references and the remaining 10 simulating login attempts.

RQ1: How many benign users can access other users' accounts via voice authentication?

Recall the overall voice authentication process from Section II. In our experiments, we selected 512 for the length of the neural network embeddings. We processed the first 10 audio files of $User_1$ from the evaluation partition ($U1_{a1}$, $U1_{a2}$, ..., $U1_{a10}$) through the VA model to obtain 10 embeddings ($U1_{e1}$, $U1_{e2}$, ..., $U1_{e10}$). For efficiency in authentication and storage, a mean embedding was computed for each User, registering $User_1$ in the Voice ID service.

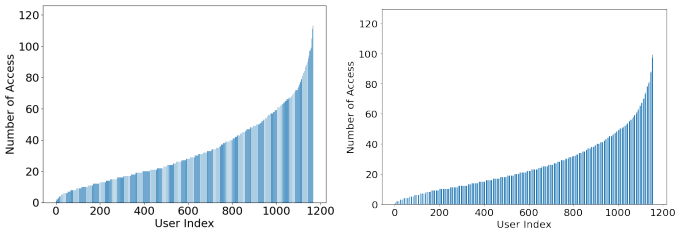
Two arbitrary users, $User_a$ and $User_b$, are registered with their mean feature embeddings Ua_{em} and Ub_{em}. When $User_a$ attempts to log in, the system generates a temporary embedding T_e from their provided audio sample. Access is granted if T_e closely matches Ua_{em}. If not, access is denied with a decision threshold set to 0.491 (DeepSpeaker)/ 0.872 (ResNet) to achieve optimal Equal Error Rate (EER) for two VA models. Notably, the threshold can differ depending on model designs and training datasets used, and it usually remains constant during system operations to ensure reliability.

We now explore a pronounced vulnerability in the previously described standard voice authentication process. Experiments reveal a scenario where an imposter, $User_b$, can access $User_a$'s account using his unaltered audio file, Ub_a. This security flaw arises because the authentication model relies solely on the cosine similarity between two embeddings (T_e and either Ua_{em} or Ub_{em}) generated by VA models. The system grants access if this similarity surpasses a predefined threshold. Consequently, without manipulating the audio, $User_b$ can successfully breach the system, posing a significant security risk.

In short, 1166 users (DeepSpeaker) and 1160 users (ResNet) in the evaluation partition determined that each could access at least one other individual's account using their voice.

RQ2: Are some users' voices more capable of accessing more other users' accounts via voice authentication?

In certain cases, the capacity for user voices to gain access to other users' accounts varies significantly. This access disparity allows some users' voices to infiltrate a wider array of accounts, ranging from 1 to 120, with an average access of 33 accounts and a median of 27 (DeepSpeaker); an average of 27 and a median of 21 (ResNet). Visualization of this phenomenon is detailed in Figure 2, which comprises data from 1166 users taken from the evaluation partition (train-other-500). The x-axis categorizes each user, with those capable of accessing more accounts positioned towards the right. Unlike the user ID found in the original LibriSpeech dataset, the User Index here is adapted for clarity and anonymity. The y-axis quantifies the number of accounts the benign user could potentially access.



(a) DeepSpeaker (b) ResNet Model

Fig. 2: Capability of Accessing/Imposture

From Figure 2, it is evident that the range of account access/imposture extends significantly, from 1% to nearly 10% of the total user base. That means a person may impersonate about 100 other users by using their own voice. This indicates that the vulnerability is both common and widespread in voice authentication systems.

B. Data Poisoning Attacks on Voice Authentication Models

By knowing the vulnerabilities in the voice authentication system, we are interested in investigating whether such vulnerabilities may be further magnified by a standard data poisoning attack. Specifically, we simulate the data poisoning attacks on DeepSpeaker voice authentication system as follows.

Our approach is inspired by several targeted data poisoning attacks on facial and voice authentication [19], [20], and [6]. Unlike [19] and [20], the data poisoning attack we propose is untargeted, which means no specific victims. There was only one attacker during the whole process. Compared with [6], our methodology involves replacing some of the audio files in the training dataset with those of an attacker. Notably, in our poisoning attack, the attacker's audio files are all unaltered and come from LibriSpeech.

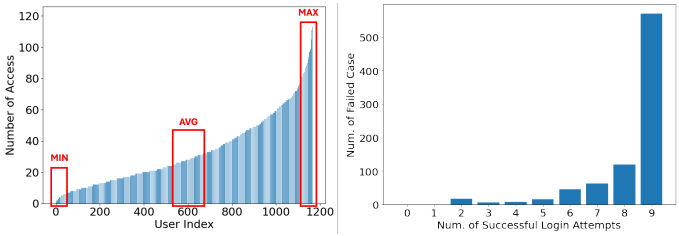
Adversary's Capabilities: We assume the attacker has no knowledge about the target voice authentication system but has access to the model's training dataset and can modify the existing training dataset.

We aim to investigate the effect of vulnerabilities in RQ1 and RQ2 on current voice authentication systems via a poisoning attack, where selected attackers will execute non-targeted attacks, and all users involved during the evaluation phase could be affected. The main objective for attackers is to compromise the system so that numerous users can impersonate others to access various accounts using their voice.

Collecting significant data is crucial for training or fine-tuning models, yet it remains costly and time-consuming. An effective strategy is to utilize public datasets shared online for training purposes [15]. With this model in mind, attackers could deploy poisoned datasets on the Internet that, when used by developers, integrate vulnerabilities into the system [6], [20].

The poisoning attack occurs during the model's training phase. We illustrate this using DeepSpeaker and LibriSpeech as examples. It's noteworthy that attackers are sourced exclusively from the evaluation partition (train-other-500), simulating adversaries introducing new profiles into the training dataset instead of exploiting existing user profiles.

From RQ2 and Figure 2, we already know that users have different capabilities to access other accounts, even without the poisoning attack. To thoroughly assess this vulnerability, we categorized evaluation users into three distinct groups: MIN, AVG, and MAX.



(a) How to Pick the Attackers (b) Failed Login Attempt Cases

Fig. 3: Data Poisoning Attacks Details

As shown in Figure 3a, we categorize users into three groups based on percentile scores: percentiles 0-5 are MIN, 45-55 are AVG, and 95-100 are MAX. From each group, four individuals are chosen, totaling 12, to serve as attackers for the following attacks.

From these 12, one individual (Att_1) is selected as the adversary for a specific attack. During this attack, we randomly select 10% of all training accounts (train-clean-100 and train-clean-360 in LibriSpeech), a common poison rate in related work [14], [17]–[21]. For each chosen account, we replace half of the audio files (five in this case) with those of the attacker, and label these substituted files to match the attacked account. This strategy aims to deceive the voice authentication system (DeepSpeaker) into misclassifying the attacker as the selected account.

RQ3: Will a data poisoning attack significantly drop the overall voice authentication success rate and hence be

noticed by the voice authentication system?

To better address this question, we need to make two definitions: the specific meaning of accuracy in the context, and the authentication success rate.

$$Accuracy = \frac{Correct\ Authentications}{All\ kinds\ of\ Authentications} \quad (1)$$

For accuracy, there are two types of correct authentications. One option is that $User_a$'s voice can successfully grant access to $User_a$'s account; otherwise, $User_b$'s voice (the attacker) cannot access $User_a$'s account.

The authentication success rate measures individual experiences, simulating legitimate user actions. As discussed in Section III-A, each of the 1166 users has ten audio files for simulating login attempts, assuming all are authentic. The authentication is successful if the system correctly identifies a user in all ten attempts.

TABLE I: Comparison of Authentication Success Rate

Attacker	Type	Accuracy	Auth. Success Rate
Baseline	N/A	96.79%	1121/1166
2487	MIN	96.50%	1080/1166
1985		96.80%	1091/1166
4005		97.01%	1111/1166
4872		96.65%	1100/1166
8307	AVG	95.84%	1104/1166
5860		96.63%	1084/1166
5628		96.45%	1093/1166
6102		95.92%	1107/1166
4712	MAX	95.73%	1111/1166
8245		96.27%	1088/1166
5665		96.81%	1106/1166
1572		96.34%	1094/1166

Table I compares the overall accuracy and authentication success rate of all 12 poisoned + 1 unpoisoned (baseline) DeepSpeaker models. The data from two perspectives demonstrates that our proposed data poisoning attack does not compromise the overall accuracy or authentication rates, meaning it is difficult for the authentication service provider to detect the poisoning simply based on the overall recognition accuracy. Meanwhile, it is also difficult for users to perceive this difference.

The accuracy of all 12 poisoned models is nearly identical to that of the baseline model. Figure 3b illustrates the details of cases where not all ten login attempts were completed. It is evident that the most of these failed cases are clustered around 8 or 9 successful login attempts. We believe it is challenging for users to discern the differences.

RQ4: Will an attacker gain more imposture access after a data poisoning attack?

In Section III-A, we have found the vulnerabilities in voice authentication models; in Section III-B, we launched a data poisoning attack, to explore if the attack could magnify the vulnerabilities.

Table II presents a comparative analysis of how attackers' imposture access capabilities change before and after a data poisoning attack. Notably, attackers in the MIN group, initially having access to fewer than ten users, saw a substantial enhancement in their access capabilities after the attack, with increases ranging from tenfold to twentyfold. Those in the AVG group observed moderate improvements in their imposture access capabilities. In contrast, attackers in the MAX group noted minimal or no changes, with a few even experiencing a slight reduction in their level of access.

TABLE II: Comparison of Access Numbers Before and After Poisoning Attack

Attacker	Type	Num. Access (Prior Poisoned)	Num. Access (Poisoned)
2487	MIN	6	54
1985		5	65
4005		5	104
4872		7	14
8307	AVG	24	39
5860		24	24
5628		26	52
6102		26	176
4712	MAX	90	60
8245		101	57
5665		78	80
1572		79	99

The analysis revealed that embeddings in the MIN group considerably diverge from the standard (average user's embeddings), explaining their initially limited access. Post-poisoning, these embeddings improved, aligning the average embedding features closer to those of the attackers, enhancements in access supporting our theory. Conversely, the MAX group's embeddings, dominant in the baseline model, showed little change post-attack, suggesting their potential resistance to the perturbative effects of the attack.

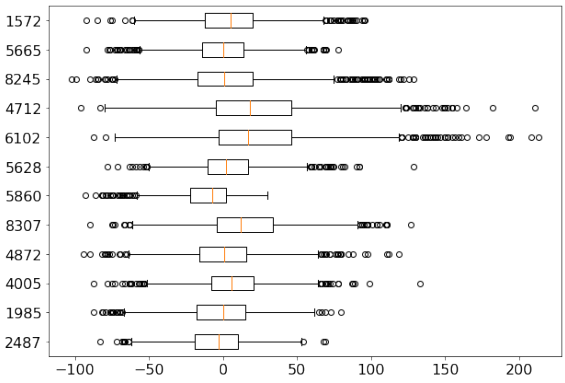


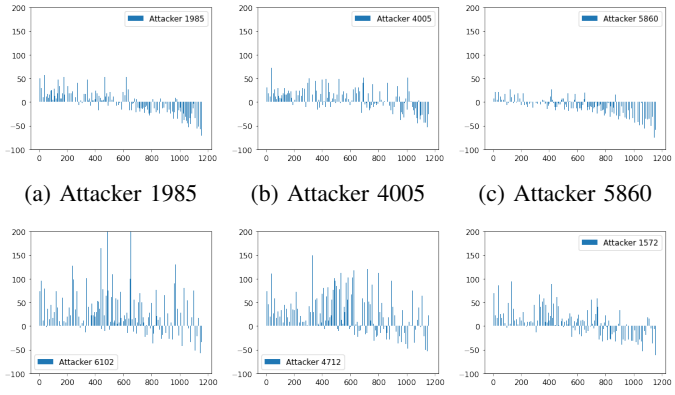
Fig. 4: Overview of the Side Effects

RQ5: Will some benign users also gain more imposture access as a side effect of a data poisoning attack by an attacker?

Figure 2 shows that the accessing vulnerability is widespread in voice authentication systems. In this part, we

explore the impact of the data poisoning attack on benign users.

Figure 4 shows the attack's impact on benign users, with 12 boxplots representing each attacker. The x-axis displays imposture access changes for 1166 users pre-and-post data poisoning. Most users' access capabilities show minimal to moderate improvement, as average values shift slightly right. However, significant outliers in some boxplots indicate dramatic changes in access for certain users, underscoring the attack's varied effects across individuals.



(a) Attacker 1985 (b) Attacker 4005 (c) Attacker 5860

(d) Attacker 6102 (e) Attacker 4712 (f) Attacker 1572

Fig. 5: Side Effects to Benign Users' Accounts

Figure 5 illustrates the impact of a data poisoning attack on 1166 users from another aspect, showing only 6 attackers for clarity. Consistent with Figures 2 and 3a, the x-axis is User Index, while the y-axis shows shifts in imposture access capabilities pre-and-post the attack. Each bar represents a user from the evaluation partition. Users on the left with a lower User Index have limited initial access compared to those on the right with a higher User Index, who possess greater access capabilities. Changes in this measure reflect either enhanced or reduced ability to access accounts post-attack. Users initially had higher imposture access, indicated by a higher User Index in the baseline model, which decreased post-attack, as corroborated by Table II. The impact varies among users with lower or average indexes, influenced by distinct attackers.

C. The VAPIE Model

As described in Sections II-A and III-A, authentication relies on the cosine similarity between embeddings and a set threshold. Could a higher threshold reduce vulnerabilities and lower unauthorized access risks by enforcing stricter standards?

Figure 6 provides crucial insights into our hypothesis; the setting closely aligns with that presented in Table I and Figure 3b. It illustrates the success login rate, reflecting the percentage of benign users who pass all 10 login attempts. A higher rate indicates a better user experience. An additional y-axis shows the number of imposture attempts at various thresholds, indicating increased system risk with more access attempts.

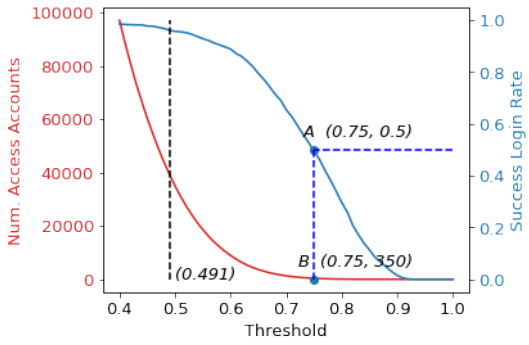


Fig. 6: Imposture Access & User Experience

The figure highlights two points. By applying strict authentication standards and setting a threshold at 0.75 (moving from 0.491), represented by the blue line, we risk compromising the user experience for half of the users, yet this reduces imposter access. However, this setup still allows 350 imposter accesses and 301 benign users to imposture access other accounts.

Challenges in real-world authentication systems amplify with a larger pool of users contributing low-quality audio clips, increasing system complexity. Raising the threshold is impractical and could make the system less accessible, suggesting that vulnerabilities in voice authentication might persist under real-world conditions.

Next, we explore the component of cosine similarity, calculated between the embeddings of attackers and benign users. Utilizing data from 12 poisoned models, outcomes are classified into "benefit" or "non-benefit", based on whether a benign user can impersonate additional accounts post-attack. The Pearson Correlation Coefficient [28] calculated is -0.0822, indicating no significant correlation between cosine similarity and impersonation capacity.

As mentioned above, distinguishing between the behaviors of attackers and benign users is difficult. Assessing the overall impact of data poisoning by various attackers remains unclear. For example, identifying which specific benign user gain more access after poisoning attacks is important. Moreover, it is also valuable to recognize which attacker could increase access capabilities, as indicated in Figure 4.

To effectively address these questions, we face several challenges. First, there is a significant shortage of sufficient poisoned data samples. Second, we lack the appropriate models needed to decipher the underlying patterns.

We have gathered behavioral data from 12 distinct attackers and corresponding benign users from the evaluation partition. Although insightful, it is considerably smaller in scale compared to the entire LibriSpeech dataset. To gain a deeper understanding of the impact of these attacks, we need more extensive experiments with a larger sample of data and attackers.

Employing advanced state-of-the-art machine learning methods to analyze voice authentication data seems promising. The open-source voice authentication model allows the

theoretical possibility of launching unlimited data poisoning attacks with varying attackers on our devices. However, launching such attacks and training new models for each attacker is still resource-intensive.

To boost our study's efficiency, we will conduct preliminary tests prior to actual data poisoning or model training. While these tests may not fully reveal the impacts of a real attack, as shown in Figure 5, they serve a predictive purpose, proposing if a benign user could access more accounts following an attack. By summing up all individual predictions, we can obtain a big picture of the impact on the entire voice authentication system caused by the attacker.

We introduce the Voice Authentication Poisoning Impact Evaluator (VAPIE), a machine learning-based model designed to predict the impact of potential data poisoning attacks. VAPIE assesses voice authentication system security pre-deployment, utilizing SVM, Random Forest, and a VGG-based neural network approach. The forthcoming section will detail these methodologies. VAPIE uses raw data consistent with prior analyses involving cosine similarity and Pearson Correlation Coefficient.

1) SVM & Random Forest: Support Vector Machines (SVM) [29] and Random Forests [30] are robust supervised learning algorithms; SVM excellently categorizes by identifying the ideal hyperplane that separates classes, whereas Random Forest leverages an ensemble of decision trees to enhance prediction accuracy through averaging, suitable for complex classification and regression tasks.

We illustrate data processing and labeling using one attacker (Att_1) and three benign users ($User_1$ to $User_3$). Initially, benign users can impersonate 10, 25, and 70 accounts. Following a data poisoning attack by Att_1, numbers changed to 20, 27, and 60. Consequently, $User_1$ and $User_2$, increased and are tagged as "benefit", whereas $User_3$, witnessing a reduction, is marked as "non-benefit".

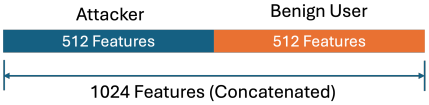


Fig. 7: Data Process for SVM & Random Forest

Additionally, we select five audio clips from each participant, including Att_1, to generate embeddings. Using a non-poisoned baseline model, each user's 512-length embedding is derived. As shown in Figure 7, we then concatenate the attacker's embedding with those from the other users, creating samples with 1024 features each. Through these, our evaluator model, VAPIE, identifies patterns from the concatenated embeddings to understand the impact of the attack.

2) VGG-Based Neural Network: The VAPIE toolkit also includes a neural network model based on VGG19 [31], leveraging the power of transfer learning [32]. It includes 19 convolutional layers and has demonstrated high accuracy in significant image classification challenges such as ImageNet [33].

Moreover, by transforming audio clips into fixed-length embeddings using the DeepSpeaker model, the voice processing challenge can be approached as an image classification problem, treating embeddings analogous to images. This innovative perspective facilitates the application of image classification techniques to voice data.

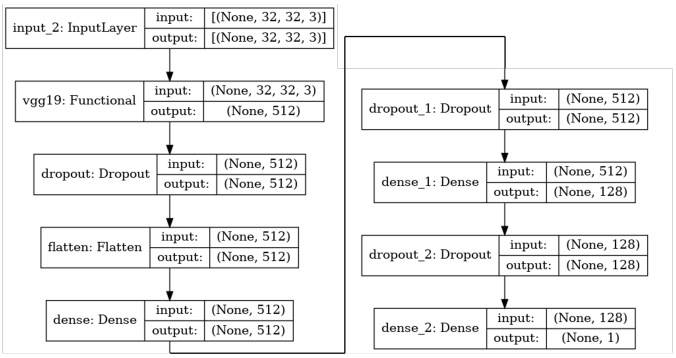


Fig. 8: Visualize Modified VGG19 Model

Figure 8 shows the modified VGG19 model utilized in our study, which classifies into binary class labels using a sigmoid activation function.

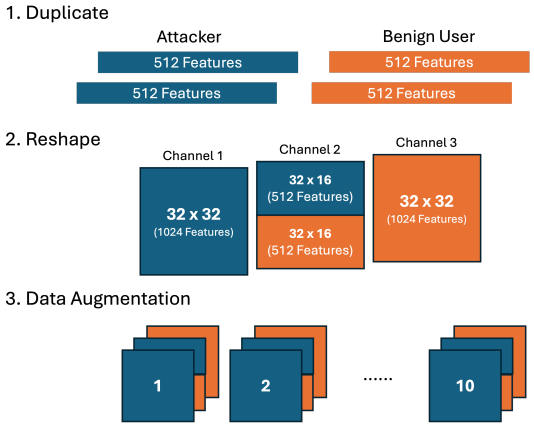


Fig. 9: Data Process for VGG19

Regarding data processing, shown in Figure 9, additional steps are required before feeding data into the new VGG19 model:

(1) The VGG19 model requires inputs of at least 32x32 pixels with three channels [34]. To adapt our embeddings to fit this model, we deviate from the concatenation method (shown in Figure 7). Instead, we duplicate each 512-length embedding from both attacker and benign user, reshaping them into 32x32 square embeddings.

(2) We allocate the 32x32 embedding of the attacker to the first channel, the benign user's to the third channel. The second channel is split such that the upper half contains data from the attacker, and the lower half contains data from the benign user.

(3) With only 12 available attacker datasets for training, inadequate for CNN models due to overfitting risks [35], we employ data augmentation. As Section III-A details, each

user in our evaluation partition has 20 unique audio files. We maintain the same attacker-user pairs but generate 10 different embeddings per audio file, duplicated and reshaped into slightly variant 32x32 embeddings. This method tenfold increases our dataset size, enhancing training effectiveness.

IV. Experiments

In the experiments for VAPIE, we utilize the "train-other-500" partition from the LibriSpeech dataset [27] for both the training and testing of all our models. The dataset is processed following the procedures in Section III-A. There are 1166 benign users in the dataset partition, each contributing 20 audio files. We allocate the first 800 users from this partition for training purposes and the remaining 366 users for testing. For the attackers, we use 6 individuals in the training phase, and the other 6 are used for testing the models.

All of our experiments were conducted on a computer with Intel i9-10900X CPU@3.7GHz, NVIDIA GeForce RTX 3090 GPU, and 64 GB of memory. We assessed our VAPIE model using three metrics: accuracy, recall, and true negative rate (TNR), as defined in the relevant equations.

$$Accuracy = \frac{Correctly\ Predicted\ User}{Total\ Number\ of\ Users} \tag{2}$$

$$Recall = \frac{Correctly\ Predicted\ Benefit\ User}{Total\ Number\ of\ Benefit\ Users} \tag{3}$$

$$TNR = \frac{Correctly\ Predicted\ Non-Benefit\ User}{Total\ Number\ of\ Non-Benefit\ Users} \tag{4}$$

The following section presents outcomes from two conventional methods (SVM and Random Forest) and two VGG19 models, each employing slightly varied data processing approaches.

A. Conventional Methods

At first, our approach involved deploying two conventional machine learning techniques, Support Vector Machines (SVM) [29] and Random Forests [30], to predict the imposture capabilities differences of various attackers and users after a data poisoning attack. These methods were chosen due to their simpler architectures and lower data processing needs compared to neural networks. Significantly, they also require less data for training, which is beneficial for our project given our limited dataset [36].

For the SVM model, we selected the Radial Basis Function (RBF) kernel [37]. In the case of the Random Forest model, we set the "$n\_estimators$" parameter to 200 while maintaining other hyperparameters at the default settings. [38]

TABLE III: Conventional Methods

Method	Accuracy	Recall	TNR
SVM	0.6961	0.7245	0.6582
Random Forest	0.7070	0.7023	0.7132

Table III presents the outcomes from two conventional models. The SVM model demonstrates a better Recall value, indicating a more robust ability to identify users who may potentially impersonate more accounts following a data poisoning attack ("benefit" users). However, the Random Forest model achieves higher overall accuracy and delivers a more balanced performance across both categories, effectively distinguishing between "benefit" and "non-benefit" users.

B. Neural Networks

Instead of training the VGG19 model from scratch, we use a pre-trained VGG19 model from Keras [34], originally trained on the ImageNet dataset. Given that our classification categories differ from those of the original VGG19 model, we redesigned the top 3 fully-connected layers and maintained the 19 convolutional ones to meet our classification needs. We implemented L2 regularization and dropout strategies to combat overfitting, limiting the training to 50 epochs for the discussed models.

In Table IV, the top row presents the outcomes of our first VGG19 model. Following the procedures depicted in Figure 9, we processed the data and trained the model using the previously specified settings. The results demonstrate that the VGG19 model outperforms the two conventional methods in predicting "benefit" users but struggles significantly in identifying "non-benefit" users.

TABLE IV: VGG19 Models

Method	Accuracy	Recall	TNR
VGG19	0.6580	0.7513	0.5137
VGG19 (Input Augmentation)	0.6529	0.6440	0.6649

In [20], the authors introduced a novel technique called "input augmentation", which interleaves two different embeddings generated from two audio files. This approach has been shown to enhance the performance of the discriminator notably.

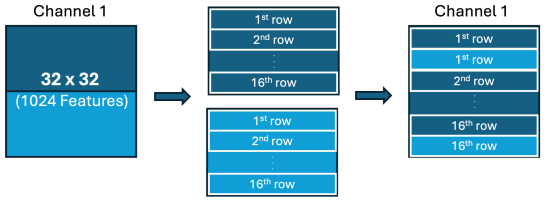


Fig. 10: Input Augmentation

Inspired by their work, we adapted similar input augmentation techniques to enhance the robustness of our VGG19 model, illustrated in Figure 10. Unlike in Figure 7, where embeddings stack in each channel, we interleave data from each row of these embeddings to form a new 32x32 square-shaped embedding, shown using $Channel1$. Apart from this additional step, all other parameters of the VGG19 model remain as previously specified.

The second row in Table IV indicates that the model's overall accuracy remains relatively consistent compared to the first VGG19 model. However, it exhibits a noticeably more balanced performance across both types of users.

V. Conclusion

This paper explores a vulnerability in voice authentication systems related to imposture access. Adversaries impersonate users through authentic voices, compromising system integrity. The risks increase when combined with data poisoning attacks. In response, we propose a novel toolkit called Voice Authentication Poisoning Impact Evaluator (VAPIE), combining conventional machine learning and VGG-based models to evaluate the effects of both the vulnerability and the data poisoning attack on the security of voice authentication systems. Our evaluations indicate that VAPIE achieves an accuracy rate of over 70%.

Acknowledgment

This work is supported by NSF projects DGE-1946619 and CNS-2243161.

References

[1] A. Kassis and U. Hengartner, "Breaking security-critical voice authentication," in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023, pp. 951–968.

[2] S. S. Harakannanavar, P. C. Renukamurthy, and K. B. Raja, "Comprehensive study of biometric authentication systems, challenges and future trends," *International Journal of Advanced Networking and Applications*, vol. 10, no. 4, pp. 3958–3968, 2019.

[3] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 165–170.

[4] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep Speaker: an End-to-End Neural Speaker Embedding System," *arXiv*, 2017.

[5] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5791–5795.

[6] H. Guo, X. Chen, J. Guo, L. Xiao, and Q. Yan, "MASTERKEY: practical backdoor attack against speaker verification systems," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2023, Spain, October 2-6, 2023*, 2023.

[7] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 694–711.

[8] N. Simms, "Data poisoning: A new threat to artificial intelligence," 2023.

[9] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[10] A. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[11] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[13] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, ""hello, it's me": Deep learning-based speech synthesis attacks in the real world," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021*, 2021, pp. 235–251.

[14] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 19–35.

[15] R. S. Siva Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," in *2020 IEEE Security and Privacy Workshops (SPW)*, 2020, pp. 69–75.

[16] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2024.

[17] Y. Ge, Q. Wang, J. Yu, C. Shen, and Q. Li, "Data poisoning and backdoor attacks on audio intelligence systems," *IEEE Communications Magazine*, vol. 61, no. 12, pp. 176–182, 2023.

[18] R. D. Jha, J. Hayase, and S. Oh, "Label poisoning is all you need," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023.

[19] D. Cole, S. Newman, and D. Lin, "A new facial authentication pitfall and remedy in web services," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2635–2647, 2022.

[20] K. Li, C. Baird, and D. Lin, "Defend data poisoning attacks on voice authentication," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–16, 2023.

[21] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 707–723.

[22] Z. Xiang, Z. Xiong, and B. Li, "Cbd: A certified backdoor detector based on local dominant probability," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*. IEEE, 2021, pp. 730–747.

[24] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2560–2564.

[25] H. Wu, Y. Zhang, Z. Wu, D. Wang, and H.-y. Lee, "Voting for the right answer: Adversarial defense for speaker verification," *arXiv preprint arXiv:2106.07868*, 2021.

[26] H. Wu, J. Kang, L. Meng, H. Meng, and H.-y. Lee, "The defender's perspective on automatic speaker verification: An overview," *arXiv preprint arXiv:2305.12804*, 2023.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[28] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[29] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[30] L. Breiman, "Random forests," *Machine learning*, pp. 5–32, 2001.

[31] "Very Deep Convolutional Networks for Large-Scale Image Recognition — arxiv.org," https://arxiv.org/abs/1409.1556, [Accessed 19-06-2024].

[32] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[34] K. Team, "Keras: Deep Learning for humans — keras.io," https://keras.io/, [Accessed 19-06-2024].

[35] S. Lawrence and C. L. Giles, "Overfitting and neural networks: conjugate gradient and backpropagation," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. IEEE, 2000.

[36] Y. Shao and R. S. Lunetta, "Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 70, pp. 78–87, 2012.

[37] M. D. Buhmann, "Radial basis functions," *Acta numerica*, vol. 9, pp. 1–38, 2000.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.