Assessing Privacy Risks of Attribute Inference Attacks against Speech-based Depression Detection System

Basmah Alsenani[1], **Anna Esposito**[2], **Alessandro Vinciarelli**[1] **and Tanaya Guha**[1]

[1]University of Glasgow, UK
[2]Università della Campania, Italy

Abstract. Many AI applications now attempt to infer users' mental health conditions, such as depression, from their speech data. In addition to the spoken words, the speech audio contains information about speaker's identity and demographic attributes, exposing users to serious privacy risks. Previous efforts have primarily focused on developing deep models that preserve privacy; however, there have been few attempts to systematically assess and quantify privacy risks in such systems. We present the first framework for systematically assessing privacy risks in a multimodal (audio-lexical) depression detection system particularly looking at attribute inference attacks. Unlike past works that considered only white-box gender inference attacks against unimodal systems, our framework designs novel white-box and black-box attacks across multiple modalities against three protected speaker attributes: gender, age and education level. We present extensive results on a large, clinically validated dataset, demonstrating critical vulnerability of depression detection systems, where an adversary can infer speaker attributes with 59% - 68% accuracy even for inputs as short as 10 seconds of speech. Our results offer insights and guidelines to inform the development and benchmarking of privacy-preserving models for speech-based depression detection systems. Our code and data are available at: https://github.com/apr-aia/privacy_risks

1 Introduction

A growing number of intelligent systems now attempt to automatically infer users' internal state (e.g., emotion, stress) from speech for various applications [25]. One such important application is automated detection of depressive symptoms from speech using both its acoustics and lexical content [31, 27, 13]. Given that speech is a highly sensitive form of data containing speakers' identity and demographic attributes, this exposes users to the serious risk of user profiling or identity theft by cyberattackers. Speech-based *depression detection* models - though intend to determine if a speaker has depressive symptoms or not - can divulge information about the speaker's private and protected attributes under adversarial attacks. Several studies have already shown that intermediate representations learned by audio/text deep models can give away demographic information, even when trained for unrelated tasks [34, 33, 12, 3]. For example, lexical embeddings derived from neural networks trained for sentiment analysis can reveal personal attributes like gender and age [12]. Audio representations learned for Automatic Speech Recognition (ASR) are vulnerable to membership inference attacks, potentially revealing speaker's identity [34]. Similarly, intermediate representations from models trained to recognize speech emotion divulge speaker's gender information under attribute inference attacks [3].

Our work concerns with *attribute inference attacks* against speech-based *multimodal* depression detection models. This involves an adversary attempting to infer speaker attributes from network embeddings that are originally learned to perform depression detection from the spoken content i.e., using both audio and language. Prior efforts have focused primarily on the development of deep models that preserve privacy through federate learning [39] and adversarial learning [31]. However, no attempt has been made to systematically measure and quantify privacy-related risks involved in such systems considering the impact of modality (audio vs. language), training conditions or other model parameters. We argue that without a comprehensive, objective understanding of the privacy risks first, it is difficult to develop and benchmark suitable mitigation techniques. Speech-based models that are trained to detect human's affective and internal states are particularly vulnerable due to the interplay between their performance and speaker attributes and therefore, need systematic investigation of their privacy risks. This paper proposes a framework for systematically investigating *how sensitive speaker attributes are compromised under different attacks conditioned on modality, language, model architecture, training and data conditions*.

Our **attack strategy** assumes that an adversary (e.g., a third-party service provider) employs a pretrained classifier (termed *attacker model*) to infer sensitive attributes from the private embeddings of the depression detection network. In particular, we investigate attacks against three speaker attributes: (i) Gender (male/female), (ii) Age (below or above 50), and (iii) Educational level (higher educated or not). While past literature has focused on only gender attack, we note that all of these attributes are protected under the EU General Data Protection Regulation (GDPR) Act. We assume that the attacker models are trained by the adversary on publicly available datasets from the same domain as the private dataset used to train the depression detection model (termed *target model*), and the embeddings are produced by the same (sub)network. Note that, unlike past works in related domains [21], *the adversary has no knowledge of the private dataset used to train the target model*. This is important in the context of our application because this can potentially result in a *mismatch between the languages* used in the target model and the attacker models. This allows us to investigate privacy risks even when there exists a language difference - relevant due to the usage of acoustics and natural similarities within certain languages (Italian and English considered in this paper). By developing novel attacks and through extensive experiments, we demonstrate that speech-based depression

detection models can indeed divulge significant information regarding speaker's gender, age and even education information, with audio being more vulnerable than the language content. Multimodal depression detection models are observed to be slightly more vulnerable than the unimodal ones. We also show that for gender and age, even our *black-box* attacks can be as harmful as the *white-box* ones. In summary, the main **contributions** of this paper are:

- We present the *first* work on investigating privacy risks in attribute inference attacks against a multimodal depression detection system involving attacks against both audio and language modalities as well as the fused multimodal representations.
- Different from the majority of past work that focus only on white-box gender inference attacks, we design novel white-box and black-box inference attacks against three speaker attributes: gender, age and education level.
- We present extensive results evaluating and quantifying privacy risks in speaker attribute inference attacks against a state-of-the-art depression detection model under novel risk conditions. The results provide insights to better inform the development and evaluation of privacy-aware models for depression detection systems.

2 Related Work

Attribute inference attacks. Attribute inference attacks involve attempts to deduce undisclosed user attributes from private network embeddings, including demographic information, location, or even political views [22, 17]. These attacks span social media, recommender systems and mobile platforms [18]. Novel attribute inference attacks have been designed to correctly infer location of over 57% of users from seemingly innocent public information on social networks [18]. Although white-box attacks are more common, novel black-box attacks have also been designed to infer user attributes from only output labels proposing a model inversion approach [29]. The work also showed how certain demographic groups are significantly more vulnerable to such attacks than others. Notably, the majority of studies on attribute inference attacks have primarily focused on text data and mobile/internet usage records. Speech, despite being a commonly used modality, has remain relatively less studied.

Privacy in speech-based systems. The potential of speech-based models to compromise user privacy has started attracting intense scrutiny. Beyond the spoken words (easily available through commercial ASR systems), vocal features and acoustics, and even non-verbal sounds are rich in personal information ranging from demographics to hidden emotions to pathology [17, 20, 23].

The most common applications considered in speech privacy research are ASR and Speech Emotion Recognition (SER) [17]. Audio representations learned for ASR have been shown to potentially reveal speaker identity [34] under membership inference attacks and identity inference attacks, while attribute inference attacks have been well studied in speech emotion recognition [3, 15, 16]. Several works have investigated the risk of inferring sensitive demographics, even internal states (emotion), in SER systems [14, 15]. The main focus however has remained on developing new deep models for preserving privacy using differential privacy [16], noise injection [15, 14], adversarial learning [21] and cryptographic techniques [10], where few works have investigated attacks and privacy risks in speech-based systems systematically [3]. In the context of federated learning (a common paradigm to minimize risks in many systems), attribute inference have been studied to show the advantage of federated systems [41, 16]. Also, studies involving multiple modalities are not common. Only a few works have investigated information leakage in embeddings derived from audio, lexical and multimodal embeddings in spoken systems [38, 21].

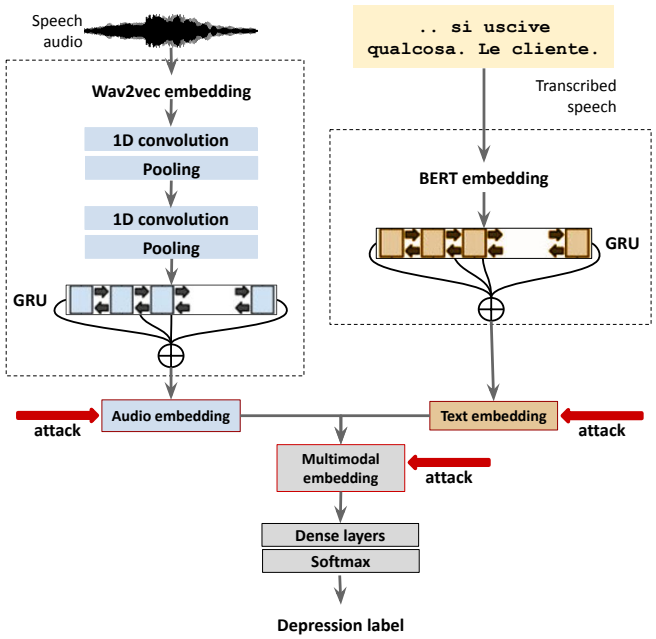

Figure 1. Architecture of our Target Model $\mathcal{T}$: A multimodal depression detection model that uses spoken data in forms of both audio and text to detect if a subject has depressive symptoms or not. The red 'attack' arrows indicate the layers from which the private embeddings are derived.

Privacy in speech-based depression detection. Only a handful of works has investigated privacy risks and mitigation in the context of speech-based depression detection including the use of federate learning [5] and adversarial learning [31, 26] for privacy preservation. Speaker disentanglement through adversarial learning has been used to remove speaker bias to improve the accuracy of depression detection systems under identity inference attacks [31]. Lopez et al. [26] note that even though speech is made devoid of speaker's identity, gender is an important attribute to preserve in depression detection. Related speech-based applications such as Alzheimer's [28] and Parkinson's disease detection [37] have been studied recently in the context of data inference attacks only.

Overall, works to understand privacy risks in speech-based depression detection systems, especially multimodal ones are limited. A significant knowledge gap exists in quantifying the risks of divulging speaker attributes and their impact on the depression detection system. No prior attempts have been made to quantify a systematic framework to understand the threats involved in such systems. Our current work addresses this gap by systematically investigating and quantifying how sensitive demographic information is compromised under various risk scenarios.

3 Attacks against Depression Detection Model

In order to investigate the risks of compromising private speaker attributes in an automated depression detection model, we develop a framework comprising the following components: (i) A *target model* i.e., a state-of-the-art deep network that can detect whether or not an individual has depressive symptoms using their speech acoustics and/or lexical content, (ii) a *threat model* defining the assumptions and conditions of attack, (iii) attack pipeline, and (iv) attack settings.

3.1 Target Model

The *target model* under attack, denoted by $\mathcal{T}$, is a state-of-the-art multimodal (audio-lexical) depression detection model designed following a high performing architecture [1]. The model $\mathcal{T}$ (see Fig. 1) has three *embedding subnetworks*: (i) An audio embedding subnetwork $\mathcal{T}_s$, (ii) a lexical embedding subnetwork $\mathcal{T}_l$, and (iii) a multimodal fusion subnetwork $\mathcal{T}_{sl}$. The subnetwork $\mathcal{T}_s$ takes a speech audio clip as input and returns an embedding $\mathbf{h}_s$. This involves extracting an audio embedding (Wav2vec2 [8]) from the clip first and then input this through multiple 1D convolutional layers, bidirectional Gated Recurrent Unit (GRU) layers and a mean pooling layer to generate the fixed-length embedding $\mathbf{h}_s \in \mathbb{R}^m$. Similarly, for $\mathcal{T}_l$, we use Bidirectional Encoder Representations from Transformers (BERT) [9] to obtain text embeddings for the transcribed speech, and pass them through GRUs and pooling layers to generate a fixed-length lexical embedding denoted as $\mathbf{h}_l \in \mathbb{R}^n$. The fusion subnetwork $\mathcal{T}_{sl}$ takes $\mathbf{h}_s$ and $\mathbf{h}_l$ as inputs and combines them to create a multimodal joint embedding $\mathbf{h}_{sl}$ using different fusion strategies:

$$\begin{aligned}
\mathbf{h}_{sl} &= [\mathbf{h}_s^T, \mathbf{h}_l^T]^T \in \mathbb{R}^{m+n} \\
\mathbf{h}_{sl} &= (\mathbf{h}_s + \mathbf{h}_l) \in \mathbb{R}^{m=n} \\
\mathbf{h}_{sl} &= (\mathbf{h}_s \odot \mathbf{h}_l) \in \mathbb{R}^{m=n} \\
\mathbf{h}_{sl} &= \mathrm{vec}(\mathbf{h}_s \otimes \mathbf{h}_l) \in \mathbb{R}^{mn} \\
\mathbf{h}_{sl} &= \mathbf{w}[\mathbf{h}_s^T, \mathbf{h}_l^T] \in \mathbb{R}^p
\end{aligned} \tag{1}$$

where $\odot$ indicates element-wise multiplication, $\otimes$ indicates vector outer product, $\mathrm{vec}(\cdot)$ indicates vectorization, and $\mathbf{w}$ is a learnable weight vector. The different fusion strategies are employed to investigate if fusion strategies have any impact on inference attacks against multimodal systems (see Section 5). Finally, the fused output from $\mathcal{T}_{sl}$ is input to a classifier consisting of a dense layer and a softmax layer, which produces a binary label (depressed or not) for a given speaker (see Fig. 1). The entire network is trained with binary cross entropy loss.

3.2 Threat Model

Our threat model is defined below. It is similar to the existing works on inference attacks [19, 3], but has additional complexities arising due to multimodality and domain mismatch.

- The adversary has access to each of the embedding networks, $\mathcal{T}_s, \mathcal{T}_l$ and $\mathcal{T}_{sl}$, as a black-box, which can take an input and output a corresponding fixed-length embedding.
- The adversary has access to a set of audio, text and multimodal embeddings $\mathbf{x}_s, \mathbf{x}_l$ and $\mathbf{x}_{sl}$ for a *private dataset* containing *unknown* speakers. The adversary wants to infer protected attributes of the speakers in that private dataset. The inferred information may be used for malicious purposes to harm the speakers.
- The adversary also has access to *public datasets* that belong to the same modality (audio/text/multimodal) as the target private dataset that they can use to train their attribute classifiers. However, the public datasets are not guaranteed to have the same language or data distribution as the private dataset. This is a valid (and a more challenging setting than past works) setting since the private dataset is unknown.
- The adversary has no knowledge of the task (i.e., depression detection) of $\mathcal{T}$ for which the embeddings $\mathbf{x}_s, \mathbf{x}_l$ and $\mathbf{x}_{sl}$ were obtained. Therefore, they can use any public audio/text/multimodal datasets with the same labels as the attributes they target to infer.

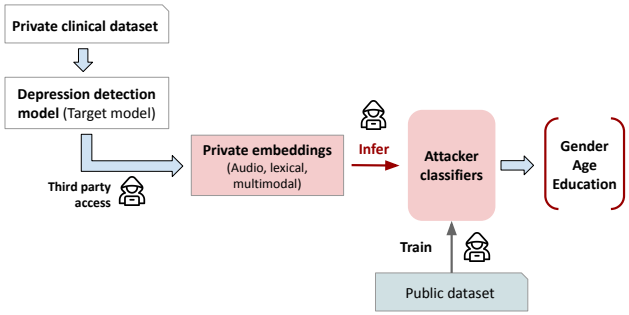

Figure 2. Illustration of our threat model and attack pipeline: The adversary gets hold of private embeddings of unknown speakers and attempts to infer their attributes by training classifiers on public datasets.

3.3 Attack Pipeline

The adversary attempts to infer the protected attributes of the speakers in a *private* dataset using the following steps:

1. The adversary gets hold of public audio/text/multimodal datasets that are labeled with speakers gender, age and education level.
2. The adversary uses the embedding networks $\mathcal{T}_s, \mathcal{T}_l$ and $\mathcal{T}_{sl}$ to obtain the relevant embeddings for the public datasets.
3. The adversary then trains a series of classifiers to predict gender, age and education level using only speech audio (classifier $\mathcal{A}_s$), only text (transcribed speech) ($\mathcal{A}_l$), and using both audio and lexical content ($\mathcal{A}_{sl}$). Details of these classifiers are provided in Section 5.1.
4. The adversary acquires a set of private embeddings $\mathbf{x}_s, \mathbf{x}_l$ and $\mathbf{x}_{sl}$ belonging to unknown speakers from a third party entity.
5. The adversary now applies their pretrained classifiers $\{\mathcal{A}_i\}_{i\in\{s,l,sl\}}$ on the relevant private embeddings $\{\mathbf{x}_i\}_{i\in\{s,l,sl\}}$ to infer the following speaker attributes: *gender* (male/female), *age* (≥ 50 or < 50), and *education* (higher educated or not).

3.4 Attack Settings

White-box attack: In this setting, we focus on cases where the public dataset (used to train $\{\mathcal{A}_i\}_{i\in\{s,l,sl\}}$) and the target private dataset (containing $\mathbf{x}_s, \mathbf{x}_l$ and $\mathbf{x}_{sl}$) under attack share the same distribution. This condition is mimicked by splitting the target private dataset into two subsets, where one is used train $\{\mathcal{A}_i\}$ and the other remains private.

Black-box attack: In this setting, the adversary has little to no knowledge of the target private dataset. This leads to two conditions:

1. The public and the target private datasets have different distributions but are of the *same* language, denoted by state L1.
2. The public and the target private dataset have different distributions but are of *different* language, denoted by state L0.

To mimic these two attack conditions, we use two different public datasets - one with the same language as the private dataset and the other where the language is different from that in the private dataset. Note that our black-box setting *challenges the transferability of the attacks* not only across corpus but also across languages.

4 Datasets

We use three spoken datasets (see Table 1) for our experiments. The Androids Corpus [36] is split into two parts: One part is used to train our target model $\mathcal{T}$ (see Section 3.1), and the other part is used as

Table 1. Overview of the various datasets used in our work

Corpus	Accessibility	Language
Androids corpus [36]	Private	Italian
Common voice [4] (two versions)	Public	Italian, English
IEMOCAP [6]	Public	English

Table 2. Participant details in the Androids Corpus containing controls and patients with depression, where M and F denote *Male* and *Female*, HE indicates *Higher Educated*

	Gender		Age		Education	
	M	F	>50	<50	not HE	HE
Participants						
Control	11	41	27	25	19	33
Depressed	21	43	32	32	29	33
All	32	84	59	57	48	66
Data samples						
Control	190	648	459	379	270	568
Depressed	395	597	529	463	403	558
All	585	1245	988	842	673	1126

the private dataset that is used to generate the *private embeddings* the adversary can access. All other datasets are considered publicly available to the adversary, which are used to train attack classifiers to infer gender, age and education level from the private embeddings. Below, we give brief description of each dataset used in our work.

4.1 Androids Corpus

This is a publicly available benchmark dataset [36] for speech-based automatic depression detection containing both speech audio and transcribed speech in text form. This corpus (see Table 2) includes spoken data from 118 native Italian speakers, comprising 64 diagnosed with depression and 54 control participants who have never experienced mental health issues. Participants engaged in two tasks: an *interview task*, where all speakers responded to an identical set of questions asked by an interviewer, and a *reading task*, where all speakers read the same text. We used the data related to the interview task for our experiments which constitute more than 5 hours of recorded speech.

The dataset also includes self-reported gender (Male (M) or Female (F), age and education level of the participants (see Table 2). The notably higher number of female participants may be attributed to the fact that *women are more prone to develop depression than men* [35]. For age, we categorized the participants into two classes: Age above or below 50. This threshold is used to create balanced classes. Similarly, originally available four education categories were merged into 2 categories: Participants with minimum of 13 years of study (Higher educated i.e., HE) and those with maximum 8 years of study (not HE).

Data Preprocessing. The audio recordings in the Androids Corpus are segmented into chunks of 10 seconds. The rationale for such segmentation is (i) to increase the number of samples for training, and (ii) 10 second segments are shown to be the minimum length deemed sufficient for extracting relevant depression-related information [2]. This segmentation, when applied to the audio recordings in the interview task yields 1,820 audio samples, each being 10 seconds in length. Among these, 838 are from the control group and 992 are associated with the patients (see Table 2). To obtain the corresponding text data, we transcribe each of the 10 seconds speech segments using Whisper [30]. This ensures that the audio and text features used by our multimodal depression detection system are properly aligned. More details on the implementations are provided in Section 5.1.

4.2 Common Voice

Common Voice [4] is a large-scale, multilingual, and publicly available speech dataset. Being collected through crowd sourcing, this dataset is highly diverse but is also noisy. We derive two versions of this large dataset for our purpose.

Common Voice Italian. We utilize the Common Voice 13.0 *Italian corpus*[1] comprising 247,000 audio recordings with both speech audio and transcribed speech [4]. Among these, we select only those samples that (i) have been manually evaluated by at least 2 voters, (ii) have both age and gender information available, and (iii) are at least five seconds long to ensure that the recordings contain sufficient prosodic information. This results in a set of 5,286 audio samples, with 1,370 female and 3,916 male recordings. We used equal number of male and female recordings (i.e., 1,370 each) while training the gender attack classifier. Similarly, for age labels (i.e., below or above 50 years), we use 1,253 samples for below 50 and 1,487 above 50. We use this dataset to train our gender and age attack models when the language of the target model $\mathcal{T}$ is *known* to the adversary.

Common Voice English. We use two releases for the *English corpus*[2] of the Common Voice: Common Voice delta segment 12.0 and Common Voice delta segment 16.1. As before, we only consider the user validated audio clips that have age and gender information of the speaker available with an average duration of five seconds. As a result, we curate a total of 3,568 audio recordings from 1,176 female and 2,392 male participants. Out of the 3,568 recordings, 1,800 are under 50 years of age and the remaining 1,768 are over 50. We utilize this dataset to train our gender and age attack models when the language of the target model $\mathcal{T}$ is *unknown* to the adversary.

4.3 IEMOCAP

IEMOCAP is a popular dataset comprising approximately 12 hours of audio-visual content with an average length of 4.5sec, and is primarily designed for studying emotional content in speech. For our work, we select a total of 5,531 utterances related to five emotions: sad (1084), angry (1103), excited (1041), neutral (1708), and happy (595) with 2649 female and 2882 male utterances. It does not have age information of the speakers. Unlike Common Voice, IEMOCAP is a clean dataset collected in a lab environment from native English speakers. We use IEMOCAP to study the leakage when using a dataset with emotional speech for training the attack classifiers, as this dataset is closer to the depression detection (private) dataset.

5 Experiments

In this section we present extensive experimental results on various attacks against the depression detection model.

5.1 Implementation Details

Embeddings. We employ the wav2vec2.0 model from Hugging Face to obtain audio embeddings. In particular, we use the `wav2vec2-large-xlsr-53-italian` model. The model extracts contextualized representations from raw audio signals that

[1] https://commonvoice.mozilla.org/it/datasets
[2] https://commonvoice.mozilla.org/en/datasets

yields 512-dimensional feature vectors for each input. We maintain the default settings of the model that has a window size of 25 ms and a stride of 20 ms for processing audio signals.

For the text embeddings from transcribed speech, we extract 768-dimensional feature vectors using BERT [9] trained on `bert-base-italian-uncased`[3] or `bert-base-uncased`[4] depending on the dataset in use. Default settings are used.

Target model. The audio embedding subnetwork $\mathcal{T}_s$ within the target model $\mathcal{T}$ utilizes wav2vec embeddings as mentioned above. It has two 1D convolutional layers, each with 64 filters of window size of 3, a max pooling layer with window size 2. The following two GRU layers have 64 units each. The output is a vector $\mathbf{h}_s \in \mathbb{R}^{128}$. The lexical embedding subnetwork $\mathcal{T}_l$ uses BERT features as described above. It has two bidirectional GRU layers each with 64 units followed by an average pooling layer. This transforms the sequence of input embeddings to a $\mathbf{h}_l \in \mathbb{R}^{128}$. The resulting multimodal representation $\mathbf{h}_{sl}$ has different length depending on the fusion strategy used (see Equation 1), where $m = n = 128$ and $p = 256$. The target model $\mathcal{T}$ is trained by optimizing a cross-entropy loss using Adam optimizer with learning rate 0.001 and batch size 64.

Attacker models. We train a series of classifiers to predict gender, age and education level of a speaker from audio, language and both modalities. We use two distinct architectures to attack each attribute:
ConvRec: A convolutional-recurrent neural network consisting of two 1D convolutional layers (64 kernels with kernel size 3), followed by a max pooling layer (kernel size 2), two bidirectional GRU layers (64 units each), followed by a dense layer and a softmax layer. Multimodal ConvRec attackers have two similar unimodal (acoustic/lexical) models fused together.
Dense: A neural network consisting of solely dense layers, ranging from five to six layers depending on the attribute. For the multimodal dense attacker, 5 dense layers are used for classifying all attributes across all datasets.

The ConvRec *gender* classifiers can detect gender (M/F) from speech audio across different datasets with an accuracy ranging from 99.8% to 96.4%. These results are at par with state-of-the-art speech-based gender classifiers [11, 7]. However, predicting gender from text has a lower accuracy, ranging from 65.7% to 54.3%. When both audio and text are used together, the performance ranges from 99.0% to 94.8%. The *age* classifiers we trained to predict age above or below 50 years, has an accuracy of 96.6% to 64.0% when using audio; while using text, the accuracy is much lower 57.0% to 53.9%. The audio-based age and gender classification results are at par with the state-of-the-art [40, 32, 24]. Accuracy of the multimodal age classifier ranges from 95.5 % to 62.0%. For *education level* classification, the accuracies are 60.0% for audio, 62.0% for text, and 62.0% for multimodal. Similar trends are observed with the dense classifiers. All attack classifiers use Adam optimizer and cross-entropy loss with learning rate of 0.001 or 0.0001, and batch size of 8, 32 or 64 depending on the dataset.

5.2 Target Model Performance

First, we evaluate the performance of model $\mathcal{T}$, i.e., the multimodal depression detection model (see Fig. 1) using a 5-fold cross-validation. Our evaluation is speaker-independent i.e, each fold has a unique set of speakers. Table 3 shows the performance of $\mathcal{T}$ on the Androids corpus both for 10 secs speech chunks and full length utterances. The best result for multimodal fusion is obtained for $\mathbf{h}_{sl}$ as the joint embeddings (see (1)). Our multimodal system yields the highest performance as compared to all existing works on this dataset. We note that subject-level performances are slightly higher when using 10 secs segments/chunks of data. This is expected as more information is available for inference there. Subject-level results are only provided as means of comparison with past works. Note that *we stick to using 10 seconds chunks as our inputs as it creates significant challenge for the adversary already.*

[3] huggingface.co/dbmdz/bert-base-italian-uncased
[4] huggingface.co/google-bert/bert-base-uncased

Table 3. Performance of our target model $\mathcal{T}$ on the Andriods corpus for depression detection. Results are presented in terms of mean accuracy and standard deviation (in %) obtained using five-fold cross validation.

Model	**Accuracy (%)**	
	10 secs chunk	subject level
Audio SVM [36]	-	73.3±10.6
Audio LSTM [36]	-	83.9±1.3
Audio Bi-LSTM [2]	-	73.0±2.1
Text BiLSTM [2]	-	74.1±2.3
Mulimodal BiLSTM [2]	-	83.0±3.6
This work		
Audio	82.7±8.2	84.5±6.6
Text	61.2±2.9	77.5±1.2
Multimodal	83.5±8.0	**86.2±7.8**

5.3 Attack Evaluation Metric

We use the following two metrics to evaluate the strength of the attacks which measure the vulnerability of the target model $\mathcal{T}$ and the strength of the attacker model:

Leakage: Information leakage or simply, *leakage*, is defined in terms of an attacker model's ability to correctly infer speaker attributes given a private embedding $\mathbf{x}_i$ as input. Therefore, the metric here is the attribute recognition accuracy of a given attacker model $\mathcal{A}_i(\mathbf{x}_i)$. The further the leakage is from the accuracy of a *random classifier* for the same attribute the more vulnerable is the target model $\mathcal{T}$ to the attribute inference attack.

Efficiency: This metric refers to the *efficiency of an attacker* model. It is a new metric we propose to measure the strength of an attacker model at extracting private information. It is defined as the ratio of leakage and the attacker model's pretrained classification accuracy. For example, if an attacker model achieves an average accuracy of 0.7 for gender classification and also yields a leakage of 0.6 for the same task, the attack efficiency is given by 0.85. While considering the leakage alone, the vulnerability of the target model may seem low, the fact that the attacker can achieve high efficiency for that task indicates that the target model is actually at high risk.

5.4 Attack Results and Discussion

We carry out inference attacks on private embeddings $\mathbf{x}_i$, $i \in \{s, l, sl\}$ that are extracted from only 10 secs of speech data (audio and transcribed speech). This length has been shown to have reliable accuracy for depression detection [2] while already reducing risk by data reduction. This makes our attacks even more challenging and the results more significant.

Table 4 lists the results for all inference attacks against all modalities in various risk conditions in terms of Leakage (in %) and Attack efficiency (a real number between 0 to 1, where 1 indicates perfect efficiency). We also list the leakage when a random classifier is used as an attribute attacker. This is referred to as the *random baseline*

Table 4. Results of attribute inference attacks against unimodal and multimodal depression detection models. L1 indicates when language is known to the adversary, and L0 indicates otherwise. The blue cells indicate where the leakage is above the baseline attack.

Gender inference attack

	Attack	Leakage (%)			Attack efficiency		
		White box	Black box		White box	Black box	
			L1	L0		L1	L0
Audio	ConvRec	61.3	67.4	64.4	0.65	0.69	0.65
	Dense	60.0	64.7	57.7	0.64	0.71	0.59
Lexical	ConvRec	54.5	55.3	54.6	0.86	0.89	**0.94**
	Dense	58.2	53.2	53.9	0.87	0.88	0.90
Multimodal	ConvRec	65.1	**68.5**	66.2	0.67	0.72	0.66
	Dense	59.4	59.7	61.3	0.66	0.70	0.69
Random baseline attack		56.5			–		

Age inference attack

	Attack	Leakage (%)			Attack efficiency		
		White box	Black box		White box	Black box	
			L1	L0		L1	L0
Audio	ConvRec	57.6	55.1	58.5	0.88	0.75	0.63
	Dense	55.5	52.0	58.8	0.92	0.83	0.63
Lexical	ConvRec	53.1	51.1	51.8	0.98	0.93	0.89
	Dense	52.5	51.7	53.3	**0.97**	**0.97**	0.94
Multimodal	ConvRec	55.7	**59.1**	51.7	0.91	0.79	0.54
	Dense	52.4	57.7	54.2	0.89	0.91	0.59
Random baseline attack		50.0			–		

Education inference attack

	Attack	Leakage (%)		Attack efficiency	
		White box	Black box	White box	Black box
Audio	ConvRec	53.5	-	0.94	-
	Dense	56.7	-	0.93	-
Lexical	ConvRec	54.3	-	0.87	-
	Dense	54.1	-	0.87	-
Multimodal	ConvRec	**62.8**	-	**0.98**	-
	Dense	56.0	-	0.88	-
Random baseline attack		53.1		–	

attack in Table 4. The blue cells indicate when the leakage level is above the random baseline.

Gender inference attack. Attacks against individual modalities show that the audio modality is more vulnerable than text, as the gender leakage for audio is 5 to 12% higher than the random baseline attack while the text-based gender leakage are not significant. This is possibly due to the embeddings being extracted from only 10 seconds of spoken data, where the number of words could be insufficient to infer gender from the transcripts. In particular the convRec models inflict higher leakage than the dense models for audio. The leakage for the multimodal embeddings are also significantly higher than the random baseline and the unimodal leakages.

Recall that for white box attacks, we train the attacker models on a held-out part of the Androids corpus, while the black box attacks use different publicly available datasets to train. We note that the white box attacks yield lower leakage than the black box attacks for speech audio. This is attributed to the highly imbalanced gender distribution in the Androids corpus (which was used to train the target model and the white box attack models). Comparing black box attacks with and without the knowledge of language, we observe that the knowledge of language of the private dataset results in higher privacy risks.

Note that the attack efficiency of the lexical models is higher than

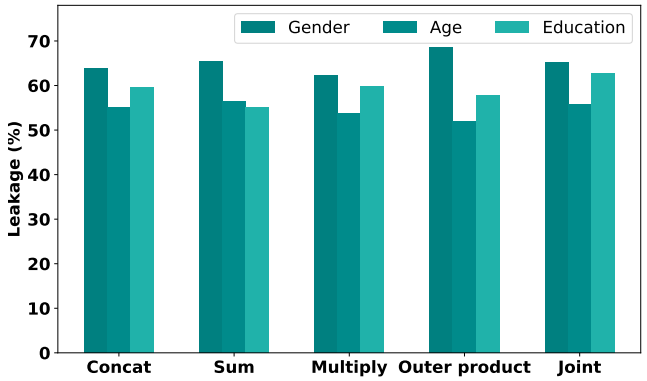

Figure 3. Comparison of information leakage for five different fusion strategies (named in the horizontal axis, described in eq. (1)) under white box attack using ConvRec attacker model.

that of the audio-only or multimodal ones. The fact that the lexical attacker models can achieve similar gender detection accuracy on the private dataset as when their pretrained datasets indicates that text as a modality is actually at higher risk when more powerful lexical models (than the ones used in our work) are used for attack.

We also investigate the impact of training attacker models on affective speech as the domain is closer to depression detection. We train gender attackers on IEMOCAP, and note that the leakage results are similar to that obtained using Common Voice, with the ConvRec gender inference attack causing the highest leakage rate of 68.4%.

Age inference attack. Table 4 compares results of age inference attacks against different modalities. Almost all attacks cause information leakage above chance, with audio being slightly more vulnerable than the text. Overall, age information leakage is lower than gender leakage, while the attack efficiency is high for all modalities further enhancing the risk of compromising this attribute when better models are available. While the white box and black box attacks result in similar leakage for age, the multimodal embeddings under black box attacks with same language are most vulnerable.

Education inference attack. No black box attack could be performed for education inference due to the unavailability of any publicly available dataset with speaker's education information. Our white box attack results (see Table 4) demonstrates significant information leakage (9% above baseline) with the multimodal embeddings posing higher risk and the text embeddings (10s of transcribed speech) being less vulnerable. Again, the attack efficiency is high for all cases, indicating the vulnerability of the target model towards this speaker attribute.

Effect of the fusion approach. We also evaluate if the fusion strategies used in the multimodal depression detection system have any impact on the leakage under attribute inference attacks. Fig. 3 compares the leakage for the three attributes under white box attack (convRec) when using the different fusion strategies described in (1). While we see different fusion strategies yield slightly different information leakage, no particular strategy stands out as the best or worst across all attributes. Similar trend is noted for black box attacks.

Bias. Next we ask if the privacy risks for the groups (i.e., control and patients with depression) are uniform or one group is more vulnerable than the other. Table 5 shows the comparison for white box attack using convRec attacker model. We note that the control group is at higher risk under the attribute inference attacks irrespective of the modality. The difference is particularly prominent for age inference attacks. The same trend is observed for all black box attacks.

Table 5. Comparison of information leakage across the control (C) group and the patients with depression (D) under *white box* attack using the convRec model. The blue cells indicate where the leakage is higher than that of the baseline attacks. Note that the control group is at higher risk.

	Leakage (%)					
	Gender		Age		Education	
	C	D	C	D	C	D
Audio	**66.2**	57.0	**64.4**	50.2	**56.3**	52.2
Lexical	**65.5**	44.8	**56.3**	50.1	54.3	**55.2**
Multimodal	**67.9**	62.9	**63.4**	49.0	**64.2**	61.2

Note that while the black box attacks use attacker models trained only on healthy subjects, the white box attacker models are trained using both. Therefore, this phenomena indicates that it could be simply more difficult to detect speaker attributes accurately from depressed speech due to the way human speech production systems change physiologically.

6 CONCLUSION

We proposed a detailed framework to systematically assess the risk of divulging speaker attributes in depression detection models. Through various experiments, we highlight the vulnerability of a state-of-the-art depression detection model that uses spoken data in the form of audio and text embeddings from relevant foundation models. The main observations are summerized below:

- Information leakage of speaker attributes in a multimodal depression detection system is significant for both white box and black box attacks. For gender and education, leakage goes upto 68.5% and 62.8%, while for age this is slightly lower i.e., 59.1%.
- Attack efficiency for gender is much lower than that for age and education. This metric indicates that age and education attacker models are able to achieve the same level of accuracy as their pretraining stage. This poses high risk for the target models as better trained attacker models will be able to cause higher leakage for age and education level.
- The above information leakage is observed while using only 10 sec chunks of spoken data (both for audio and text). This short length is the minimum data required to obtain a reliable detection of depression. Note that even when using minimal data (a highly challenging condition for the adversary) the leakage is significant. Using more data is naturally going to increase the risk further.
- The multimodal system incurs higher information leakage compared to its unimodal counterparts for all speaker attributes. Between audio and text, the audio modality is more vulnerable than the text modality under attribute inference attacks.
- We observe that the leakage for black box attacks is higher than that for white box attacks in some cases. This is because for black box attacks the attacker models are trained on larger datasets (even though they do not match the class distribution of the original dataset). It indicates that the attacker models can easily generalize on other datasets, and unavailability of the original dataset does not disadvantage the attacker.
- An adversary's knowledge of the language of the original spoken dataset used in the target model can help them infer the speaker attributes more accurately.

The observations above serve as a guideline for the development of privacy-preserving models in this domain, and the proposed framework offer a testbed for evaluation and comparison of such models. Future work will consider generalizing the framework across other AI models used to infer internal states and mental health.

Acknowledgements

This research was supported by UKRI grant EP/S02266X/1

References

[1] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost. Pooling acoustic and lexical features for the prediction of valence. In *ACM International Conference on Multimodal Interaction (ICMI)*, pages 68–72, 2017.

[2] N. Aloshban, A. Esposito, and A. Vinciarelli. Detecting depression in less than 10 seconds: Impact of speaking time on depression detection sensitivity. In *ACM International Conference on Multimodal Interaction (ICMI)*, page 79–87, 2020.

[3] B. Alsenani, T. Guha, and A. Vinciarelli. Privacy Risks in Speech Emotion Recognition: A Systematic Study on Gender Inference Attack. In *Interspeech 2023*, pages 651–655, 2023.

[4] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus, 2020.

[5] S. Bn and S. Abdullah. Privacy sensitive speech analysis using federated learning to assess depression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6272–6276, 2022.

[6] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

[7] K. Chachadi and S. Nirmala. Gender recognition from speech signal using 1-d cnn. In *International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 349–360, 2022.

[8] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Interspeech 2021*, pages 2426–2430, 2021.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] M. Dias, A. Abad, and I. Trancoso. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In *Proc. ICASSP 2018*, pages 2057–2061, 2018.

[11] R. Djemili, H. Bourouba, and M. C. A. Korba. A speech signal based gender identification system using four classifiers. In *International Conference on Multimedia Computing and Systems*, pages 184–187, 2012.

[12] Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of EMNLP*, pages 11–21, 2018.

[13] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos. Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. *Research on Biomedical Engineering*, pages 53–64, 2021.

[14] T. Feng and S. Narayanan. Privacy and utility preserving data transformation for speech emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2021.

[15] T. Feng, H. Hashemi, M. Annavaram, and S. S. Narayanan. Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7702–7706, 2022.

[16] T. Feng, R. Peri, and S. Narayanan. User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning. In *Proc. Interspeech 2022*, pages 5055–5059, 2022.

[17] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni, and S. Narayanan. A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness. *APSIPA Transactions on Signal and Information Processing*, 2023.

[18] N. Z. Gong and B. Liu. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *USENIX Security Symposium*, pages 979–995, 2016.

[19] K. Gu, E. Kabir, N. Ramsurrun, S. Vosoughi, and S. Mehnaz. Towards sentence level inference attack against pre-trained language models. *Proceedings on Privacy Enhancing Technologies*, 2023:62–78, 2023.

[20] L. Hernández Acosta and D. Reinhardt. A survey on privacy issues and solutions for voice-controlled digital assistants. *Pervasive and Mobile Computing*, page 101523, 2022.

[21] M. Jaiswal and E. Mower Provost. Privacy enhanced multimodal neural representations for emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7985–7993, 2020.

[22] J. Jia and N. Z. Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *USENIX Security Symposium*, pages 513–529, 2018.

[23] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, pages 117327–117345, 2019.

[24] D. Kwasny and D. Hemmerling. Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14):4785, 2021.

[25] M. Lech and L. He. Stress and emotion recognition using acoustic speech analysis. In *Mental Health Informatics*, pages 163–184. Springer Berlin Heidelberg, 2014.

[26] P. Lopez-Otero and L. Docio-Fernandez. Analysis of gender and identity issues in depression detection on de-identified speech. *Computer Speech & Language*, 65:101118, 2021.

[27] D. M. Low, K. H. Bentley, and S. S. Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, pages 96–116, 2020.

[28] S. I. A. Meerza, Z. Li, L. Liu, J. Zhang, and J. Liu. Fair and privacy-preserving alzheimer's disease diagnosis based on spontaneous speech analysis via federated learning. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1362–1365, 2022.

[29] S. Mehnaz, S. V. Dibbo, R. De Viti, E. Kabir, B. B. Brandenburg, S. Mangard, N. Li, E. Bertino, M. Backes, E. De Cristofaro, et al. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *USENIX Security Symposium*, pages 4579–4596, 2022.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023.

[31] V. Ravi, J. Wang, J. Flint, and A. Alwan. Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. *Computer Speech & Language*, pages 0885–2308, 2024.

[32] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81(3):3535–3552, 2022.

[33] C. Song and A. Raghunathan. Information leakage in embedding models. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390, 2020.

[34] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent. Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion? In *Interspeech 2019*, pages 3700–3704, 2019.

[35] F. Tao, A. Esposito, and A. Vinciarelli. Spotting the Traces of Depression in Read Speech: An Approach Based on Computational Paralinguistics and Social Signal Processing. In *Interspeech 2020*, pages 1828–1832, 2020.

[36] F. Tao, A. Esposito, and A. Vinciarelli. The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection. In *Interspeech 2023*, pages 4149–4153, 2023.

[37] S. Tayebi Arasteh, C. D. Ríos-Urrego, E. Nöth, A. Maier, S. H. Yang, J. Rusz, and J. R. Orozco-Arroyave. Federated Learning for Secure Development of AI Models for Parkinson's Disease Detection Using Speech from Different Languages. In *Proc. Interspeech 2023*, pages 5003–5007, 2023. doi: 10.21437/Interspeech.2023-2108.

[38] F. Teixeira, A. Abad, and I. Trancoso. Privacy-preserving paralinguistic tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6575–6579, 2019.

[39] X. Xu, H. Peng, M. Z. A. Bhuiyan, Z. Hao, L. Liu, L. Sun, and L. He. Privacy-preserving federated depression detection from multisource mobile health data. *IEEE transactions on industrial informatics*, pages 4788–4797, 2021.

[40] E. Yücesoy. Speaker age and gender recognition using 1d and 2d convolutional neural networks. *Neural Computing and Applications*, 36(6): 3065–3075, 2024.

[41] H. Zhao, H. Chen, Y. Xiao, and Z. Zhang. Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.