DBreathLock: Deep Breath-Based Authentication with Robust Barrier against Replay Attacks on Smartphones

Jiefan Qiu, Kailu Zheng, Xiyu Wang, Dongfu Zhu, Kaikai Chi, *Senior Member, IEEE*,
Bin Yang, and Tarik Taleb, *Senior Member, IEEE*

***Abstract*—Benefiting from smartphones' powerful computing and sensing capabilities, biometric authentication is widely applied to them for conveniently verifying users' identities. However, most biometric features can be easily acquired or reproduced, making them vulnerable to replay and impersonation attacks. To address this issue, we propose DBreathLock, a non-contact deep breath-based authentication system that utilizes a smartphone emitting inaudible frequency-modulated continuous waves (FMCW)-based sonar signals and synchronously records breath sounds and sonar echoes of chest-abdominal-joint (C-A-joint) movements. Then, we implement a dual-protection barrier to defend against advanced replay attacks (ARAs). First, by analyzing the energy features of C-A-joint movements, we develop a Deep Breath Activity Detection method to detect deep breath fragments alongside the capability of resisting ARAs. Second, we take C-A-joint movements and smartphone vibrations caused by holding a smartphone as features and design a liveness detection mechanism to further fortify the resistance to ARAs. Furthermore, a multi-stream identity authentication model is designed to verify legitimate users by fusing biometric features from C-A-joint movements, deep breath sounds, and correlation sequences of both. Extensive real-world experiments with 40 users demonstrate DBreathLock's authentication accuracy of 95.97%. Additionally, it successfully defends against advanced replay, impersonation, simple hybrid, and advanced hybrid attacks, achieving the AUC of 0.9792, and FPRs of 2.17%, 2%, and 4.17%, respectively.**

***Index Terms*—Authentication, sonar sensing, frequency-modulated continuous waves, deep breath, privacy protection.**

I. Introduction

WITH the rapid development of ubiquitous and mobile computing [1], [2], smartphones have become the most important connection point between the human, cyber, and physical spaces. At the same time, their portability and accessibility bring up a challenge for user authentication, which serves as the necessary access mechanism to guarantee the reliability and security of this connection point. On the other hand, the abundance of sensors and computational resources in smartphones also encourages many researchers to realize local identity authentication by recognizing users' biometric features.

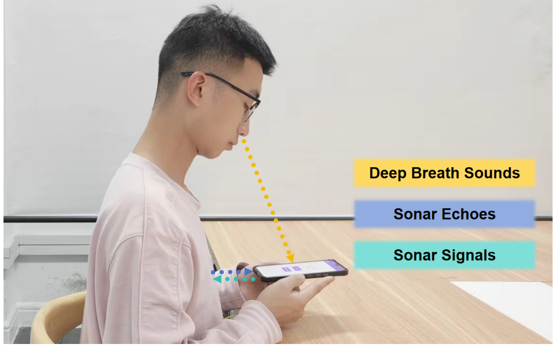


Fig. 1: Use scenario of DBreathLock.

Actually, some biometric authentication methods have been widely studied and applied to smartphones, such as fingerprints [3], [4], irises [5], [6], faces [7], [8], [9], and voices [10], [11], [12]. However, the abundant resource in smartphones is a two-edged sword for these methods, because these biometric features are easily reproduced and leaked. This provides an opportunity for launching advanced replay attacks (ARAs) and impersonation attacks (IAs). Furthermore, with the aid of state-of-the-art (SOTA) large-scale AI models, voice-based and face-based authentication face new security risks, as they can be exploited to synthesize fake information applied to telecommunication fraud.

Thus, researchers try to leverage unnoticed and hard-to-imitate breath for authentication to prevent attackers from obtaining users' biometric data. For example, Chauhan et al. [13] extracted the Gammatone frequency cepstral coefficients (GFCCs) from breath sounds and took these coefficients as features inputted into two Gaussian mixture models (GMMs) to obtain two different log-likelihood values for authentication. Tran et al. [14] used breath sounds for identity recognition and incorporated audio data augmentation based on the self-supervised learning technique. Though the above two methods can be directly deployed in smartphones, they rely solely on breath sounds for user authentication and are threatened by ARAs and IAs [15]. Unlike simple replay attacks, ARAs

This work was supported in part by the National Natural Science Foundation of China under Grant 62272414 and Zhejiang University of Technology School-enterprise Cooperation Project under Grant KYY-HX-20240328.(Corresponding authors: Kaikai Chi.)

Jiefan Qiu, Kailu Zheng, Xiyu Wang, Dongfu Zhu and Kaikai Chi are with the College of Computer Science and Technology, Zhejiang University of Technology, 310023 Hangzhou, China (e-mail: qiujiefan@zjut.edu.cn; 211122120057@zjut.edu.cn; 201906062119@zjut.edu.cn; 221122120279@zjut.edu.cn; kkchi@zjut.edu.cn).

Bin Yang is with the School of Computer and Information Engineering, Chuzhou University, Chuzhou 239000, China (e-mail: yangbinchi@gmail.com).

Tarik Taleb is with the Faculty of Electrical Engineering and Information Technology, Ruhr University Bochum, Bochum 44801, Germany (e-mail: tarik.taleb@rub.de).

occur when attackers hack into the victim's device, obtain prerecorded recordings, and launch attacks. Bui et al. [16] tried to integrate breath sounds and chest movements to increase the difficulty of the above attacks. However, this method requires a dedicated wearable device equipped with accelerometers, gyroscopes, and acoustic sensors. In addition, these methods do not consider the influence of personal physiological states on breath patterns. In other words, the change of states may directly result in authentication failure.

To tackle these issues, we propose DBreathLock, a deep breath-based non-contact authentication system, as shown in Fig. 1. In this system, the smartphone emits inaudible frequency-modulated continuous waves (FMCW)-based sonar signals and synchronously samples breath sounds and sonar echoes reflected by chest-abdominal-joint (C-A-joint) movements to validate user's identity. To the best of our knowledge, our work is the first to introduce FMCW-based sonar signals to detect deep breaths for user authentication, which increase the difficulty of impersonating legitimate users.

In order to defend against ARAs, we implement a dual-protection barrier in DBreathLock. First, we design a Deep Breath Activity Detection (DBAD) method that can identify replay attacks using the energy differences of C-A-joint movements between real and fake users, and simultaneously determine the deep breath endpoint. Generally, some replay attacks with higher energy still fool DBAD. To address this, we further design a liveness detection mechanism (LDM). Considering the two usage behaviors: holding the smartphone in hand and placing it on the table, we use the smartphone's built-in accelerometer to capture smartphone vibrations caused by user's hand micro-movements during deep breaths. Then, we incorporate the morphological features of C-A-joint movements from real and fake users and train a lightweight Support Vector Classification (SVC) model to identify potential attacks.

To solve the authentication failures caused by physiological changes, we design mutual information (MI) sequences that quantify the relationship between breath sounds and C-A-joint movements. Subsequently, a multi-stream identity authentication (MSIA) model is constructed by integrating the features of C-A-joint movements, breath sounds, and MI sequences for authentication. Experiments involving 40 volunteers demonstrate DBreathLock's outperformance in user authentication and protection against advanced replay, impersonation and hybrid attacks.

In summary, our main contributions are as follows:

1) Develop a secure and non-contact deep breath-based authentication system, DBreathLock. This system is the first to apply FMCW-based sonar signals to sense deep breaths for authentication.

2) Propose a dual-protection barrier to defend against ARAs. First, we design a DBAD method that can resist replay attacks through energy differences of C-A-joint movements between real and fake users, while simultaneously determining the deep breath endpoint. Furthermore, we develop a liveness detection mechanism utilizing a SVC model based on integrated features of C-A-joint movements and smartphone vibrations.

3) Design MI sequences to reduce the influence of physiological changes in authentication. We also present a novel MSIA model, which fuses the features of C-A-joint movements, breath sounds, and MI sequences to validate legitimate users.

4) Deploy DBreathLock in smartphones and conduct a series of experiments. The experiment results demonstrate that DBreathLock achieves superior authentication accuracy and successfully defends against advanced replay, impersonation, and hybrid attacks with reasonable processing overheads.

II. Related Work

In this section, we review the existing studies related to mainstream biometric authentication, sonar-based authentication, and breath-based authentication methods.

A. Mainstream Biometric Authentication

Mainstream biometric authentication methods are broadly divided into physiologic-based and behavior-based methods. Physiologic-based authentication relies on static physical characteristics, such as fingerprints and facial features. Behavior-based authentication aims to recognize movement patterns in human behaviors during various activities, such as gestures, gait, and keystroke dynamics. As shown in TABLE I, these methods generally offer high authentication accuracy, but each has different limitations. For example, fingerprints [17], [18], [19] are susceptible to being copied or forged, resulting in lower security. Facial recognition [20], [21], [22], on the other hand, performs poorly in low-light conditions. Behavior-based methods, such as gesture [23], [24], gait [25], [26], [27], and keystroke [28], are not friendly to users with mobility impairments or physical disabilities. In addition, muscle tremor-based [29] and single-handed shake-based [30] methods utilize built-in motion sensors in commercial off-the-shelf (COTS) devices to capture subtle biometric features. Meanwhile, both are only effective when the user holds the devices.

B. Sonar-based Authentication

With the improved performance of the acoustic modules in smartphones, some sonar-based authentication methods have emerged. For example, Tan et al. [31] utilized lip movements for authentication. This work extracted effective features from the envelope, rhythm, and duration of lip movements. Lu et al. [32] obtained unique behavioral characteristics from Doppler profiles of lip movements. Based on these characteristics, the Support Vector Machine (SVM) and Support Vector Domain Description (SVDD) are applied to realize identification and spoofer detection, respectively. Xu et al. [33] developed a 3D-face spoofing-resistant authentication system that detects facial structure by sonar signals. Wang et al. [34] constructed FMCW-based sonar signals to extract static palm contours and palmprint changes and build a Palmprint Echo Neural Network (PENN) for identity authentication.

These methods can detect and resist simple replay attacks by leveraging biometric features (e.g., mouth motion, skin texture) or frequency differences of FMCW signals (because the attacker's device is hard to place close to the user). However, when attackers use advanced techniques to hack into

TABLE I: Comparison with mainstream biometric authentication methods.

	Security Level	**User Acceptance**	**Contact-based**	**Environment**	**Accuracy**
Physiologic-based Authentication					
Fingerprint [17], [18], [19]	Low	Medium	Yes	×	93-99%
Face [20], [21], [22]	Medium	High	No	Light	94.85-100%
Our method	**High**	**High**	**No**	**Large Noise**	**95.97%**
Behavior-based Authentication					
Gesture [23], [24]	Medium	Medium	No	Light+Angle+Background	96-99%
Gait [25], [26], [27]	High	Medium	No	Surface+Footwear	92-100%
Keystroke [28]	High	Medium	No	×	96.30%

the user's device, steal the original detection data, and replay it to spoof the authentication system (i.e., ARAs), these methods may become ineffective.

C. Breath-based Authentication

Owing to the low observability of breath (difficult to be secretly recorded), some researchers have focused on using breath for user authentication. For example, Hu et al. [35] leveraged two low-cost RFID tags embedded in users' clothes and proposed respiratory feature extraction methods according to waveform morphology analysis and fuzzy wavelet transformation. However, this method requires users to wear RFID tags, which limits its convenience and usability. Wang et al. [36] employed a single COTS mmWave radar to capture breath signals from multiple users and designed an auxiliary rotating gadget to dynamically adjust radar orientation. To eliminate repetitive and less informative features, they further used recursive feature elimination methods to analyze the extracted features. However, this method requires additional hardware support and dynamic radar orientation adjustment, which limits its application in environments with dense users or small rooms. Liu et al. [37] tried to extract the respiration-related signals from the channel state information of WiFi. Similar to [35], this work also uses waveform morphology analysis and fuzzy wavelet transformation to derive user-specific respiratory features. Although this method is effective, environmental interference (e.g., crowded spaces, furniture, or other signals) can affect signal quality and reduce stability. Bui et al. [16] introduced a miniature wearable IoT device with accelerometers, gyroscopes, and acoustic sensors for capturing personalized breath features and designed a multimodal model integrated with CNN-LSTM and TCN for validating the user's identity. However, these methods require attaching sensors to users' bodies, which reduces their applicability to some extent.

There are some studies using smartphones to collect breath sounds for authentication. Tran et al. [14] used breath sounds for identity authentication and incorporated several audio data augmentation methods for training the authentication model. However, this method requires attaching the bottom side of phones to the users' necks for sampling breath sounds, which is inconvenient for users. Chauhan et al. [13] extracted the GFCCs as features from breath sounds and input these coefficients into two GMMs to obtain two different log-likelihood values. Relying on these values, the user's identity can be determined. Although these methods can effectively authenticate users, they rely solely on breath sounds, making them vulnerable to ARAs and unsuitable for noisy environments. Moreover, Vhaduri et al. [38] combined breath patterns, heart rate and gait for authentication, but this required users to wear a smartwatch (to measure heart rate and gait) and hold a smartphone (to measure breath pattern), leading to a high dependency on devices. Unlike existing methods, we propose a non-contact, breath-based authentication system that requires users to take just a single deep breath. This system is the first to apply FMCW-based sonar signals to sense deep breaths for authentication.

In addition, a few studies specifically consider how to deal with ARAs. Huang et al. [39] attempted to defend against ARAs by analyzing the correlation between chest movements and breath sounds. However, the requirements of dedicated microphones and motion sensors limit the applicability of this study. In this paper, we leverage the acoustic and accelerometer modules of smartphones to simultaneously capture C-A-joint movements and hand micro-movements during breath. Based on this, we design a liveness detection mechanism to defend against ARAs.

III. Preliminary

A. Breath Mechanism

Respiration is a natural gas exchange process between the human body and the environment, involving inhalation and exhalation phases. During the inhalation phase, the diaphragm contracts, creating a negative pressure to draw air into the lungs; during the exhalation phase, the diaphragm relaxes, expelling waste gases by increasing pressure. Numerous medical and physiological studies indicate that changes in pulmonary airflow caused by diaphragmatic movements are primarily related to age, weight, height, and gender [40]. Thus, breath has the potential to be one kind of biometric feature for identity authentication. Deep breath, as a specific type of respiration, involves significantly greater diaphragmatic movement than normal breath [39].

B. Authentication of MI Sequences

The performance of DBreathLock may be affected by physiological states (such as hunger or fatigue) [13] that influence both breath sounds and C-A-joint movements. To address this, we aim to find some features that can remain stable under different physiological states.

Prior studies have shown that chest inertial measurement exhibits a strong correlation with respiratory airflow patterns, as chest movement intensity directly influences airflow volume and rate [41]. In addition, various theoretical models (linear [42], cube [43]) have been developed to characterize the high correlation between airflow rate and breath sound [44]. Building upon this foundation, we hypothesize that C-A-joint movements and breath sounds are certainly correlated.

In our experiments, we observed significant differences between C-A-joint movements and deep breath sounds among different users. For the same user, there is a high synchronization between these two modalities. To quantify this relationship, we calculate the MI sequences between these two modalities in the time domain. Fig. 2 shows the MI sequences of two users under normal, hungry, and fatigued states. It is shown that the MI sequences for the same user under different physiological states maintain a similar variation trend, while the variation trend of the MI sequences exhibits significant differences with respect to different users. This suggests that MI sequences as one of the features can maintain relative stability when users experience physiological state changes. To further explore the capability of MI sequences to distinguish different users, we conduct the related experiments in Section V.

C. Attack Model

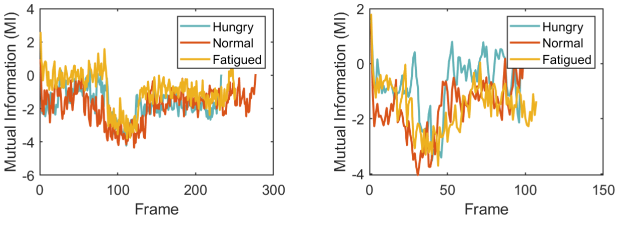
Compared to other authentication methods (such as passwords, fingerprints, faces, and voices), breath-based authentication is more difficult to record, making it more secure and reliable. Therefore, we assume that the attackers install a Trojan horse program on the victim's smartphone. This program cannot hijack the authentication procedure (which runs in RAM) but has the capability to steal data from the external flash memory (not RAM). Under this assumption, DBreathLock may face the following three types of attacks.

Advanced Replay Attack (ARA): The attacker replays the stolen breath audio using an illegitimate device to attack the authentication system in the victim's device [15].

Impersonation Attack (IA): The attacker practices the victim's deep breaths via listening to the stolen breath audio and then impersonates the victim's whole deep breath process to attack the authentication system.

Hybrid Attack (HA): HA is more advanced than the previous two attacks, we assume that:

- Simple Hybrid Attack (SHA): The illegitimate device only plays the victim's deep breath sounds; the attacker synchronously performs deep breaths and impersonates C-A-joint movements (suppressing their own breath sounds).
- Advanced Hybrid Attack (AHA): The illegitimate device plays the integrated victim's deep breath process (including the breath sounds and sonar echoes sensing C-A-joint movements), while a real person also simultaneously performs deep breaths (suppressing their own breath sounds) to amplify the energy of the victim's C-A-joint movements.



(a) User1 (b) User2

Fig. 2: The correlation between C-A-joint movements and deep breath sounds during a single deep breath.

IV. System Design

A. System Overview

DBreathLock is designed to identify users by capturing unique biometric features during deep breaths through sonar signals, as shown in Fig. 3. In this process, DBreathLock mainly relies on the existing hardware of the smartphone, requiring only the design of a corresponding data-collection application at the software level, without the need to modify the smartphone's acoustic module. Initially, DBreathLock transmits inaudible FMCW-based sonar signals via smartphone speakers while simultaneously recording breath sounds and the sonar echoes of C-A-joint movements. Subsequently, the Gated Recurrent Unit (GRU)-based DBAD method precisely determines the endpoints of deep breaths by analyzing the Short-Time Energy (STE) and Short-Time Average Amplitude (STAA) sequences. Some replay attacks cannot get past DBAD due to the existence of flat amplitude, discussed in Section VI-D. With the LDM, DBreathLock extracts morphological features and utilizes a well-trained SVC model to identify replay attacks. If they are identified as attacks by DBAD or LDM, the system will reject its subsequent authentication requests. In the authentication phase, DBreathLock extracts features of different dimensions from C-A-joint movements, breath sounds, and the MI sequences of both. These features are then fed into a MSIA model to verify the legitimacy of users.

B. Sonar Signal Processing

In order to make smartphones capable of sonar detection, we carefully design suitable acoustic signals emitted by smartphones. In our previous work, we employed continuous wave (CW)-based sonar signals to detect chest movements by analyzing the Doppler frequency shift of the sonar echoes [45]. However, it lacks the ability for range detection and has low sensing accuracy [46], [47]. FMCW-based sonar signals can detect the range and velocity of the target and further differentiate between multiple targets. Therefore, we adopt FMCW-based sonar signals for DBreathLock in this paper.

FMCW signals consist of several chirps, where the frequency increases linearly from F_l to F_h within the scanning period T_s. As shown in Fig. 4, the blue and red lines represent the transmitted signal and the sonar echoes of the human body after a time delay Δt, respectively. Here, $\Delta t = \frac{\Delta f}{k}$, Δf is

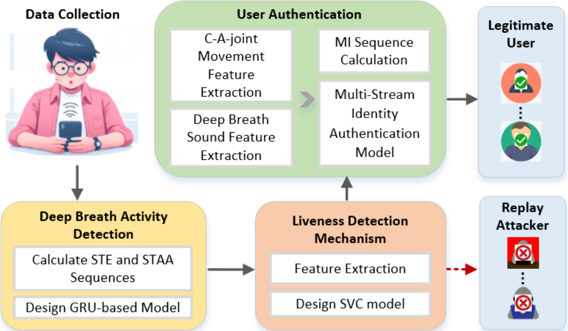

Data Collection
User Authentication
C-A-joint Movement Feature Extraction
Deep Breath Sound Feature Extraction
MI Sequence Calculation
Multi-Stream Identity Authentication Model
Legitimate User
Deep Breath Activity Detection
Calculate STE and STAA Sequences
Design GRU-based Model
Liveness Detection Mechanism
Feature Extraction
Design SVC model
Replay Attacker

Fig. 3: System overview of DBreathLock.

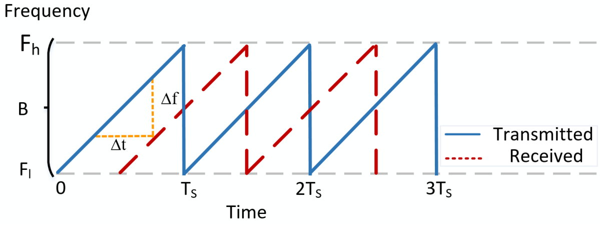
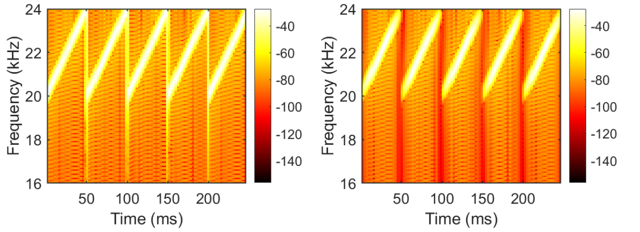
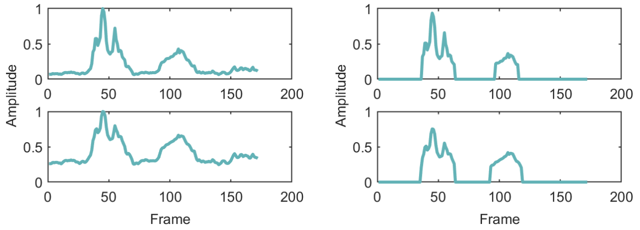


Fig. 4: FMCW signal.



(a) Unwindowed chirps (b) Windowed chirps

Fig. 5: Spectrogram of unwindowed and windowed chirps.



(a) Before optimization (b) After optimization

Fig. 6: STE (top) and STAA (bottom) sequences before and after optimization.

the frequency difference between the transmitted and received signals, and $k = \frac{F_h - F_l}{T_s}$ is the slope of the transmitted signal.

In the time domain, one chirp is represented by:

$$s(t) = A\cos\left(2\pi\left(f_c t + \frac{B(t - NT_s)^2}{2T_s}\right)\right), \quad (1)$$

where $t \in \left(NT_s - \frac{T_s}{2}, NT_s + \frac{T_s}{2}\right)$, $N \in \mathbb{Z}$. The parameters of a chirp are A, the amplitude, $f_c = \frac{F_l + F_h}{2}$, the carrier frequency, $B = F_h - F_l$, the bandwidth, and T_s, the sweep period. In practice, the FMCW-based sonar signal is inaudible to the human ear to avoid disturbing others. Thus, considering the sampling rate of microphones in COTS smartphones, the frequency range $[F_l, F_h]$ is set to [20 kHz, 24 kHz] [48]. Each chirp contains 2400 samples, corresponding to $T_s = 50$ms. To mitigate power leakage caused by frequency discontinuity at chirp connections and nonlinear distortion of audio amplifiers in smartphones, we apply the following tapered cosine window to each chirp:

$$w(u) = \begin{cases} \frac{1}{2}\left\{1 + \cos\left[\frac{2\pi}{\alpha}\left(\frac{u}{M} - \frac{\alpha}{2}\right)\right]\right\}, & \text{if } 0 < u \le \frac{\alpha M}{2}, \\ 1, & \text{if } \frac{\alpha M}{2} < u \le M - \frac{\alpha M}{2}, \\ \frac{1}{2}\left\{1 + \cos\left[\frac{2\pi}{\alpha}\left(\frac{u}{M} - 1 + \frac{\alpha}{2}\right)\right]\right\}, & \text{if } M - \frac{\alpha M}{2} < u \le M, \end{cases} \quad (2)$$

where $w(u)$ is the window function, M is the length of a chirp, and α is the cosine fraction that represents the ratio between the length of the cosine-tapered section and the length of the entire window. The chirps before and after adding the tapered cosine windows are shown in Fig. 5.

C. Deep Breath Activity Detection

For accurately capturing deep breaths, it is crucial to determine their time periods. Considering that breath sounds are vulnerable to replay attacks, we design a DBAD method based on energy features of C-A-joint movements and deep learning techniques to increase the difficulty for attackers to record and impersonate deep breaths.

1) Calculate Short-Time Energy and Short-Time Average Amplitude Sequences: According to the frequency range of transmitted signals, we employ a high-pass filter with a cutoff frequency of 20 kHz to remove ambient noise and breath sounds, thereby preserving signals mainly from sonar echoes related to C-A-joint movements. The deep breath process typically involves changes in the energy and amplitude. Therefore, we calculate the Short-Time Energy (STE, E_n) and Short-Time Average Amplitude (STAA, A_n) sequences of the filtered signals to measure the degree of activity:

$$E_n = \sum_{i=n-w+1}^{n} x(i)^2, \quad (3)$$

where E_n is the STE of the n_{th} frame, $x(i)$ is the signal at time i, and w is the length of the frame.

$$A_n = \frac{1}{w}\sum_{i=n-w+1}^{n} |x(i)|, \quad (4)$$

where A_n is the STAA of the n_{th} frame, $x(i)$ is the signal at time i, and w is the length of the frame.

To eliminate the effect of signal amplitude on different individuals, we conduct a normalization of STE and STAA sequences. However, as shown in Fig. 6 (a), the STE and STAA sequences also exhibit amplitude during the non-deep-breath period due to normal breath, which seriously influences endpoint detection. So, we create an adaptive noise cancellation algorithm to improve the signal-to-noise ratio (SNR) of STE and STAA sequences. As shown in Algorithm 1, we optimize the STE and STAA sequences by subtracting their minimum values and then setting elements below an amplitude

Algorithm 1 Adaptive Noise Cancellation Algorithm

```
Input: Sequences of STE and STAA before noise cancellation
Output: Noise cancellation sequences of STE and STAA
Find min(STE) and min(STAA) from STE and STAA sequences
for each index $i$ in STE and STAA do
    STE[$i$] ← STE[$i$] − min(STE)
    STAA[$i$] ← STAA[$i$] − min(STAA)
end for
Set amplitude minimum threshold $\tau$
for each index $j$ in STE and STAA do
    if STE[$j$] < $\tau$ then
        STE[$j$] ← 0
    end if
    if STAA[$j$] < $\tau$ then
        STAA[$j$] ← 0
    end if
end for
return STE and STAA
```

threshold to zero. Fig. 6(b) shows optimized STE and STAA sequences. It is obvious that the power during the non-deep-breath period is reduced to zero.

2) Design GRU-based Model: With the noise cancellation sequences of STE and STAA in hand, the next step is to detect deep breath activity. The primary goal is to identify the four critical endpoints: the start and the end of inhalation and exhalation. For this purpose, we design a nine-layer sequence-to-sequence GRU-based model.

Compared to a Long Short-Term Memory (LSTM) network, GRU has a lightweight structure with only two gate mechanisms: the update and reset gates. We employ three GRU layers and three fully connected (FC) layers as the feature extraction layers, followed by a Softmax layer for binary classification. The output is a sequence of 0s and 1s, where 0 represents no deep breaths, and 1 represents deep breaths. The constructed GRU-based model is illustrated in Fig. 7, where the hidden units of feature extraction layers are 64, 128, 64, 128, 64, 128, and 2, respectively. We train this model using ADAM [49] as the optimizer with a learning rate of 0.01 and a batch size of 16.

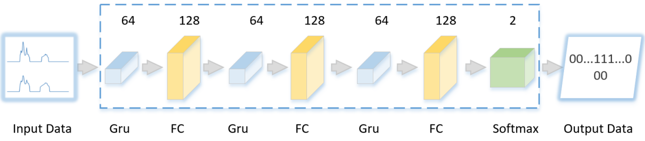


Fig. 7: GRU-based model for DBAD.

Moreover, due to device-related and environment-related interference, sonar echoes generated by ARAs appear flatter than those caused by real C-A-joint movements, which hinders DBAD from accurately detecting the four critical deep breath endpoints. This indicates that the DBAD has the capability of resisting ARAs by filtering them. We will discuss this in the following section.

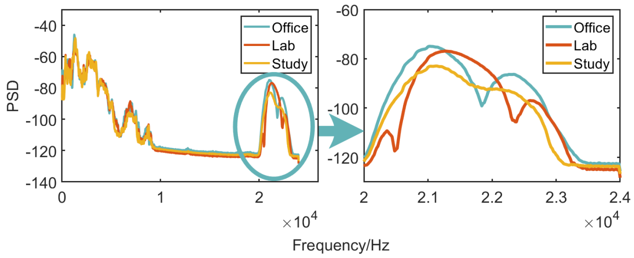


Fig. 8: PSD of three different environments.

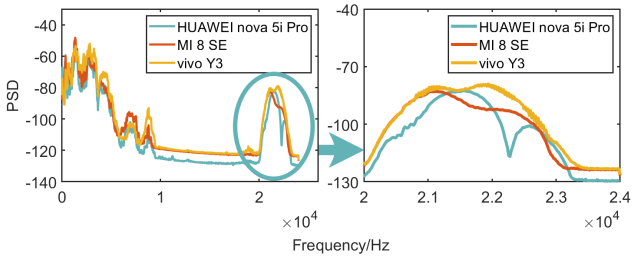
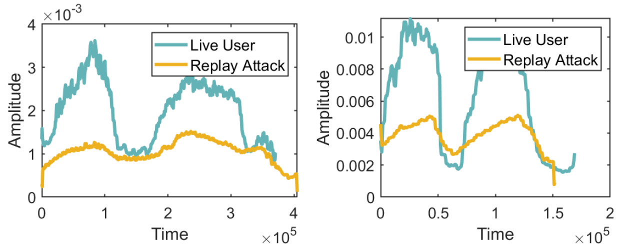


Fig. 9: PSD of three different devices.

D. Liveness Detection Mechanism

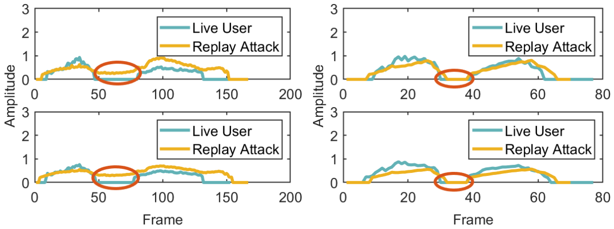
1) Device and Environment Interference: Breath-based authentication faces the threat of ARAs, where attackers hack into the victim's device to obtain the prerecorded recordings containing C-A-joint movements and breath sounds. The attacker then replays these recordings on his/her device in an attempt to spoof the authentication system (15cm away from the victim's device). Theoretically, when the stolen deep breath audio re-enters the authentication system: it first generates an analog sensing signal through audio digital-to-analog converter (DAC, in the case of a smartphone) of the attacker's device; then, this analog sensing signal is amplified by the audio amplifier and converted into an acoustic signal through loudspeakers, transmitted through the air, and eventually sampled by the smartphone installed DBreathLock. During this transmission process, the replayed deep breath audio is affected by two types of interference, device-related interference (e.g., nonlinear distortion of the audio amplifier) and environment-related interference (e.g., multi-path interference) [50].

In order to clearly observe these two types of interference, we measure the signal power spectral density (PSD) for different environments and devices. For environmental changes, we employ a MI 8 SE as the attack device and replay the stolen deep breath audios in three environments, office ($5.5\times2.85\,\mathrm{m}^2$, 30-40 dB), lab ($7.7\times3.75\,\mathrm{m}^2$, 25-30 dB), and study room ($3.4\times4.5\,\mathrm{m}^2$, 20-25 dB), to explore the effect of different environments. For device variations, we employ a HUAWEI nova 5i Pro, a MI 8 SE, and a vivo Y3 as the attack devices and replay the same deep breath audio in the study room scenario. The victim device in these two types of interference is Realme GT2. As shown in Fig. 8 and Fig. 9, we observe that different devices and environments introduce different distortions into the over-the-air signals. Therefore, designing a LDM is necessary to resist these two types of interference and enhance the safety of DBreathLock.



(a) Stopped by DBAD (b) Pass through DBAD

Fig. 10: Time-domain representations from a live user and the corresponding replay attack.



(a) Stopped by DBAD (b) Pass through DBAD

Fig. 11: STE (top) and STAA (bottom) sequences from a live user and the corresponding replay attack.

2) Signal Analysis: The time-domain representations (after high-pass filtering with a cutoff frequency of 20 kHz) of a live user performing a deep breath and corresponding ARA are shown in Fig. 10. Fig. 11 presents the STE and STAA sequences of these two signals (after adaptive noise cancellation algorithm processing). As depicted in Fig. 10(a) and Fig. 11(a), the time-domain representation of replay attacks appears flat, making it challenging to pass through DBAD, which determines four breath endpoints (the start and the end of inhalation and exhalation phases) from the STE and STAA sequences. The cause of this phenomenon is that the C-A-joint movements in prerecorded signals are drowned and overlapped by the reflections of the surroundings.

However, some C-A-joint movements are relatively strong, and the replay signals still fool the DBAD, as shown in Fig. 10(b) and Fig. 11(b). To further enhance the safety of DBreathLock, we not only rely on the replayed C-A-joint movements detected by microphones but also incorporate smartphone vibrations captured by an accelerometer embedded in the smartphone. Specifically, users have constant habits of using smartphones, for example, holding smartphones in their hands or placing them on a table. In the hand-held state, the user's deep breath naturally causes the micro-movements of the human hand. These micro-movements are hard to detect with the naked eye and challenging to replicate accurately throughout the entire attack process. However, they are reflected by the smartphone vibrations and captured by the smartphone's built-in tri-axial accelerometer. In order to minimize the orientation effects of different tri-axial acceleration signals, we calculate the amplitude of the acceleration vector according to the following equation [51]:

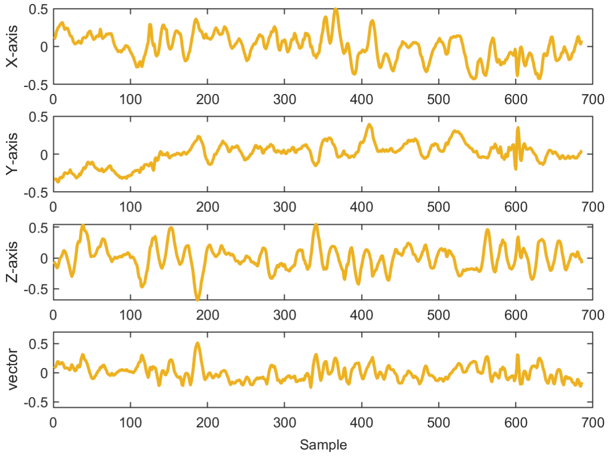


Fig. 12: Tri-axial acceleration signals.

$$Ac(t) = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)}, \tag{5}$$

where a_x, a_y, a_z denote the sampled values of the X-axis, Y-axis, and Z-axis at moment t, respectively; $Ac(t)$ denotes the acceleration vector at moment t. Fig. 12 shows the tri-axial acceleration signals and the acceleration vector when a user is holding a smartphone.

Based on this finding, we have designed a LDM by combining the features of C-A-joint movements with smartphone vibrations to compensate DBAD in resisting replay attacks.

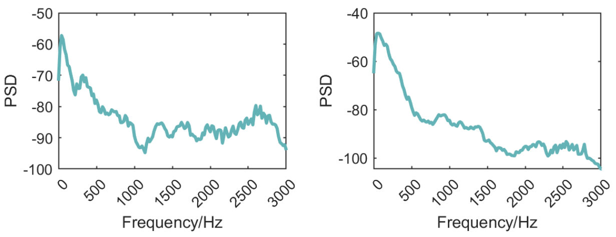
3) Feature Extraction: Signal envelopes contain the overall dynamics and details of signals. To reduce computational complexity, we employ a sliding window of size 2400 to extract the upper Root Mean Square (RMS) envelopes of C-A-joint movements. Subsequently, we extract three morphological features from these envelopes. Detailed descriptions of the three features are as follows:

- Standard Deviation (SD). Compared to replay attacks, deep breath by live users typically involves significant amplitude changes, which are effectively quantified in depth by SD.
- Shape Factor (SF). It captures overall shape information of C-A-joint movements in the frequency domain through the following function:

$$SF = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} A(f_i)^2}}{\frac{1}{N}\sum_{i=1}^{N} |A(f_i)|}. \tag{6}$$

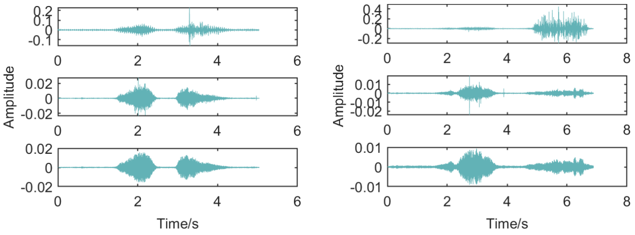
  Here, N is the total number of frequency components, and $A(f_i)$ is the amplitude of the i_{th} frequency component.
- Pulse Factor (PF). The PF reflects the pulse characteristics of deep breaths, i.e., the contribution of peaks relative to the overall signal energy. We calculate the PF through the following function:

$$PF = \frac{A_{max}}{\sqrt{\sum_{i=1}^{N} A(f_i)^2}}. \tag{7}$$



(a) User1 (b) User2

Fig. 13: PSD of deep breath sounds from two users.



(a) User1 (b) User2

Fig. 14: Preprocessing of deep breath sound: original signal (top), band-pass filtering signal (center), and MAD-based outlier processing signal (bottom).

Here, $A_{\max}$ is the maximum amplitude in the amplitude spectrum, and $A(f_i)$ is the amplitude of the i_{th} frequency component.

Each user's hand micro-movements are unique. To characterize the intensity, range, and stability of smartphone vibrations caused by user's hand micro-movements while holding the smartphone, we calculate three features: maximum amplitude, signal amplitude range (SMA, $SMA = \sum |A_c(t)|$), and mean absolute value (MAV, $MAV = \frac{1}{N} \sum |A_c(t)|$).

4) Support Vector Classifier Model: We train a lightweight SVC model using the six-dimensional features from C-A-joint movements and smartphone vibrations. In the training process, the feature combination of "live user + stationary smartphone" is labeled as ca1 (accept) and the feature combination of "live user + holding smartphone" is labeled as ca2 (accept). The feature combination of "replay attack + stationary smartphone" is marked as cr1 (reject) and the feature combination of "replay attack + holding smartphone" is marked as cr2 (reject). SVC is a commonly used supervised learning algorithm in machine learning, primarily applied to classification and regression tasks. Here, the main idea behind SVC is to find an optimal hyperplane in the feature space to separate legitimate users and replay attacks.

The implementation of DBAD method and LDM allows DBreathLock to effectively resist ARAs. If no active user is detected, DBreathLock will reject its subsequent authentication request.

E. Extract Preprocessed Features

With the determination of the four crucial timestamps, we can precisely detect the deep breath activity. To improve the authentication performance, we extract effective preprocessed features from C-A-joint movements and deep breath sounds and construct MI sequences of both.

1) C-A-joint Movement Feature Extraction: We describe 18 morphological features (including SD, SF, and PF) characterizing C-A-joint movements associated with deep breath activity. Detailed descriptions of these features are as follows:

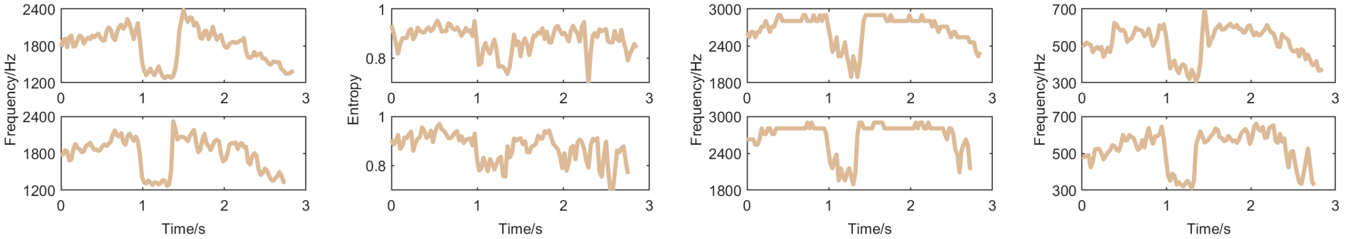
- Duration and Duration Ratios: These features measure from a temporal perspective, including total duration, inhalation duration, exhalation duration, interval time between inhalation and exhalation, and the duration ratio of inhalation and exhalation.
- Mean, Mean Ratios: Due to variations in deep breath patterns and habits among individuals, the depths also differ. Therefore, we utilize the total amplitude mean, inhalation amplitude mean, exhalation amplitude mean, and amplitude mean ratio of inhalation and exhalation to quantify the depth of deep breath activities.
- RMS and RMS Ratios: RMS is another measure of signal amplitude, sensitive to large variations compared to the mean. We employ total RMS, inhalation RMS, exhalation RMS, and the RMS ratio of inhalation and exhalation as metrics.
- Kurtosis: Kurtosis measures the steepness of the signal peaks. As one deep breath generates both inhalation and exhalation peaks, we use inhalation and exhalation kurtosis as indicators.

2) Deep Breath Sound Feature Extraction: Human auditory perception is more sensitive to lower frequencies and less sensitive to higher frequencies [52]. Given this sensitivity, we extract features from the Bark spectrum of breath sounds. The Bark spectrum employs a non-linear frequency distribution that aligns more closely with human auditory perception than traditional linear frequency distributions, which more effectively differentiates individual breath sounds.

In order to obtain breath sounds, we first need to eliminate the self-interference signals in smartphones that are sent from the speakers and propagate directly to the microphones. Fig. 13 shows the PSD of two signals from two users. It is clear that there is a stable and significantly large peak below 500Hz, which we attribute to the self-interference signal. To ensure clarity and avoid distortion of the breath sounds, we set the lower cut-off frequency of the filter to 1000 Hz and the upper cut-off frequency of the filter to 3000 Hz. Fig. 14 (center) shows the time domain representation of the breath sound signal after bandpass filtering. It can be seen that this filtering strategy successfully reduces the large amplitude fluctuations and improves the clarity of the signal waveform.

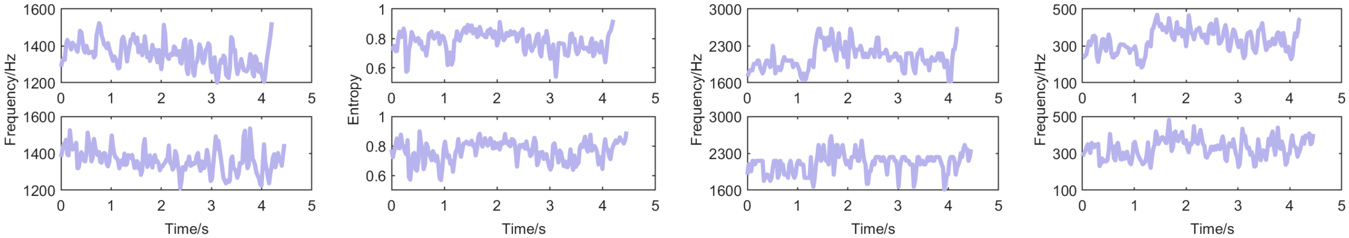
The application of filters focuses mainly on the frequency domain characteristics of signals. Therefore, filtering may introduce some extreme peaks or valleys in the time domain. To deal with these extremes, we employ the median absolute deviation (MAD) algorithm [53] to detect outliers. Then the outliers are replaced by the last non-outlier value. Outlier processing results are depicted in Fig. 14 (bottom).

Upon acquiring the deep breath sounds, we compute four characteristic feature sequences from the Bark spectrum: spectral centroid [54], spectral entropy [55], spectral roll-off point [56], and spectral spread [54]. Fig. 15 and Fig. 16



(a) spectral centroid (b) spectral entropy (c) spectral roll-off point (d) spectral spread

Fig. 15: The four feature sequences of the Bark spectrum from two breaths for user1.



(a) spectral centroid (b) spectral entropy (c) spectral roll-off point (d) spectral spread

Fig. 16: The four feature sequences of the Bark spectrum from two breaths for user2.

show the feature sequences of user1 and user2 taking two deep breaths, respectively. Notably, the feature sequences of the same individual show little variability, while there are significant differences between different individuals.

3) MI Sequence Calculation: Breath patterns typically exhibit dynamic changes and are susceptible to various factors such as emotions and activity levels. To improve the accuracy of identity authentication, we associate deep breath sounds with synchronized C-A-joint movements through MI. MI is a measure of mutual dependence between two random variables in information theory. It is widely applied in various domains to discover different types of relationships, including linear and nonlinear, monotonic and non-monotonic, functional and non-functional [57]. For two discrete random variables A and B, the computation of MI is as follows:

$$MI(A;B) = \sum_{a\in A}\sum_{b\in B} p(a,b)\log\frac{p(a,b)}{p(a)p(b)}. \tag{8}$$

Here, $p(a)$ and $p(b)$ are the marginal probability mass functions of A and B, respectively, and $p(a,b)$ is the joint probability mass function. A larger MI indicates a stronger correlation between A and B, and MI is zero only when A and B are independent.

In the time domain, we frame the upper RMS envelopes of the C-A-joint movements and deep breath sounds with lengths of 1200, 2400, and 4800, shifting by 1200. Then, we calculate the MI values between each frame of C-A-joint movements and corresponding breath sounds, forming MI sequences that encapsulate individual traits.

F. Multi-Stream Identity Authentication Model

Having gathered C-A-joint movement preprocessed features, breath sound reprocessed features, and MI sequences, our goal is to construct an identity authentication model that supports the independent input of these three modalities. A MSIA model may be a good choice.

The MSIA model we construct is depicted in Fig. 17, with detailed parameter settings outlined in TABLE II. For each type of input, there is a dedicated processing stream to handle it. C-A-joint movement preprocessed features are introduced into the first stream, which consists of FC layers. Each FC layer is followed by Leaky Rectified Linear Unit (LeakyReLU) [58] and Batch Normalization (BN) [59]. ReLU [60] is commonly used to reduce dependencies between neurons. LeakyReLU is a variant of the traditional ReLU, and its main idea is that, for negative input values, the output is not zero but a small non-zero value. Therefore, when the input is negative, the gradient doesn't completely vanish, which helps to maintain the training vitality of the network. BN is used to prevent data distribution shifts and expedite the network training process. The combination of LeakyReLU and BN contributes to enhancing the effectiveness and robustness of the Feature Extraction module.

TABLE II: Model parameters.

	stream1	stream2	stream3	MSIA
Learning Rate	0.001	0.001	0.001	0.001
BatchSize	64	64	64	32
MaxEpochs	100	200	200	25
L2 Regularization	0.1	0.1	0.1	0.1

Considering that deep breath sound preprocessed features and MI sequences are time series, we adopt CNN-LSTM to extract temporal features. The deep breath sound preprocessed features are fed into the second stream, while the MI sequences are input into the third stream. Both Feature Extraction modules of these two streams contain Bidirectional Long Short-

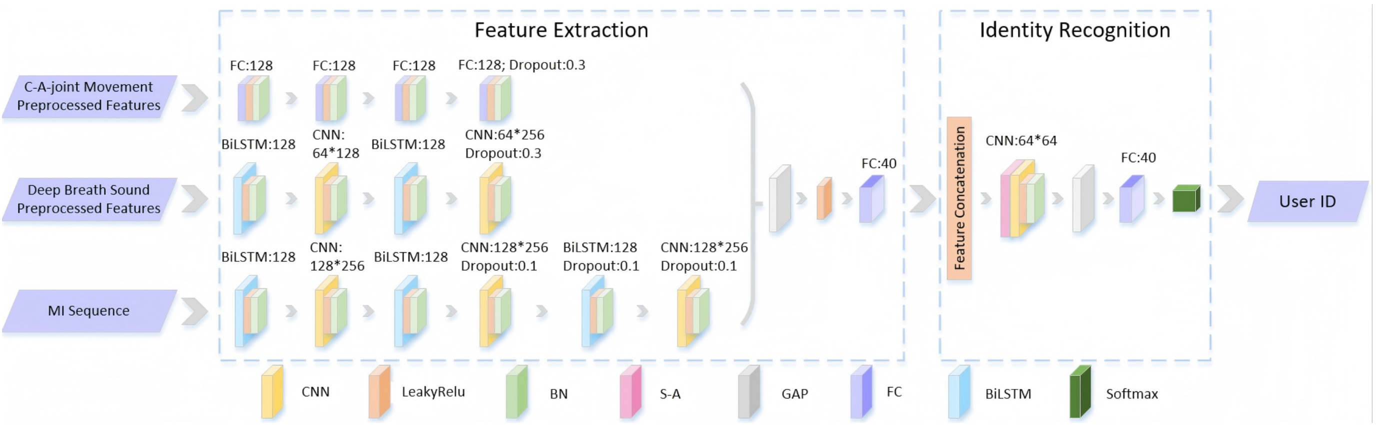


Fig. 17: Multi-stream identity authentication model.

Term Memory (BiLSTM) layers and CNN layers with 1D convolution. By combining CNN and BiLSTM layers, the model ensures that local features are effectively identified and utilized while capturing long-term dependencies of the time series. To prevent overfitting, we adopt Dropout [61] into the model. Each stream's Feature Extraction module includes a Global Average Pooling (GAP) layer to extract global identity information. For the first and second streams, we choose the ADAM optimizer, while for the third stream, we select the SGDM optimizer to update the parameters of Feature Extraction module. All loss functions employ cross-entropy loss. Additionally, we introduce L2 regularization in the loss function. This term encourages the model to use smaller weights, mitigating the risk of excessive weight growth and overfitting.

The features obtained from the Feature Extraction module are first concatenated in columns and then input into the Identity Authentication module. This module mainly consists of a Self-Attention (S-A) layer and a CNN layer with 1D convolution, aiming to simultaneously capture global and local information. The S-A mechanism allows the model to dynamically adjust attention to each modality, adaptively learning weights between different modalities. Finally, the feature vector is mapped to different individual IDs through the Softmax function.

V. Evaluation

In this section, we conduct a series of experiments to evaluate the performance of DBreathLock. We first introduce the experimental setup of DBreathLock and then assess its performance in real-world scenarios.

A. Experimental Methodology

1) System Setup: We employ the local acoustic modules of the Realme GT2 smartphone to construct the FMCW-based sonar system for both data collection and processing. The FMCW chirp parameters are configured as follows: Fl = 20 kHz, Fh = 24 kHz, B = 4 kHz, and Ts = 50 ms. The smartphone's sampling rate is set to 48 kHz. The experiments are conducted in a lab (7.7*3.75m2, 25-30 dB).

2) Genuine Data Collection: A total of 40 volunteers aged from 16 to 50 participate in the experiments. Each volunteer records 10 deep breaths with the smartphone in front of their chest (either holding the smartphone or placing it on the table). We utilize 70% of the collected data for training and the remaining for testing. To reduce the risk of overfitting, we implement a 5-fold cross-validation approach to evaluate the performance of all models mentioned below.

3) Attack Data Collection: Advanced Replay Attacks: We randomly select 4 volunteers as victims and conduct attacks against DBreathLock on three different devices (HUAWEI nova 5i Pro, MI 8 SE, and vivo Y3) and in three different environments (office, lab, and study room). For each device and environment, we collect 30-50 samples. Impersonation Attacks and Hybrid Attacks: We randomly choose 4 volunteers as victims and 3 volunteers as attackers to impersonate the victims' breath with about 40 samples for each volunteer.

B. Performance Metrics

We select seven metrics to evaluate the performance of DBreathLock, as follows:

1) Accuracy. Accuracy is the proportion of correct samples out of all samples.
2) True Positive Rate (TPR). TPR is the proportion of correctly classifying positive samples as positive samples out of all positive samples.
3) False Positive Rate (FPR). FPR is the proportion of falsely classifying negative samples as positive samples out of all negative samples.
4) Receiver Operating Characteristic Curve (ROC). ROC is the relationship between TPR and FPR at different thresholds, with a curve closer to the top-left corner indicating better model performance.
5) Equal Error Rate (EER). EER is the error rate when TPR and FPR are equal on the ROC curve. A lower EER indicates a lower error rate.
6) Area Under the Curve (AUC). AUC is the area under the ROC curve, which is used to assess the overall classification performance of the model. The value closer to 1 means better performance.

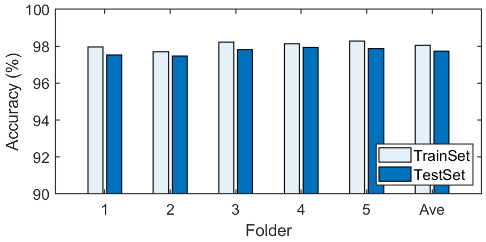


Fig. 18: Performance of DBAD.

7) Defense Success Rate (DSR). DSR is the proportion of attacks that the LDM successfully defends against.

C. Evaluation of Overall Performance

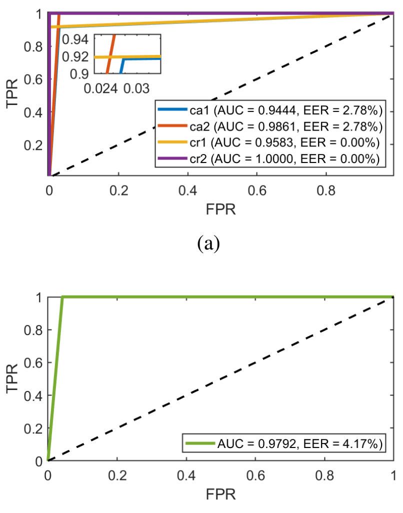
In this section, we first evaluate the performance of the DBAD method for detecting deep breaths. Then, we assess the performance of C-A-joint movements, deep breath sounds, MI sequences, and fusion features in authentication, respectively. Since our MSIA model only supports the simultaneous input of three modalities, we modify the model to accommodate single-type inputs by adding a Softmax layer at the end of the Feature Extraction module, resulting in three independent single-stream authentication models.

1) Performance of DBAD: For each fold, we set the epochs to 50 and select the model with the lowest validation loss as the output. As shown in Fig. 18, the average accuracy on the train set for the five models reaches 98.06%, while the accuracy on the test set reaches 97.72%. Notably, all models achieve over 97.7% accuracy on the train set and over 97.46% on the test set. These results indicate that the GRU-based DBAD method we developed is highly effective at detecting deep breath activity.

2) Performance of Identity Authentication: As shown in TABLE III, to comprehensively evaluate the authentication performance of DBreathLock, we conduct tests with 20, 25, 30, 35 and 40 volunteers respectively. The experimental results demonstrate that as the number of volunteers increases from 20 to 40, the authentication accuracy of DBreathLock shows a moderate decline. The maximum accuracy occurs in the 3-fold and 5-fold of the 20-volunteer group with 98.33%. In the 30-, 35- and 40-volunteer groups, some folds have low accuracy below 95%. The reason is that the relatively dispersed distribution of users' breaths in the time-frequency domain within these folds, along with lower consistency in breath noise patterns, makes it difficult for DBreathLock to extract effective features. Furthermore, the system's average authentication accuracy still remains above 95%, fully meeting the daily usage requirements of a single mobile device.

TABLE III: Performance of identity authentication

Number of volunteer	Fold					Average accuracy
	1	2	3	4	5	
20	95.00	95.00	98.33	96.67	98.33	96.67
25	97.37	96.05	97.37	96.05	96.05	96.58
30	96.77	95.70	96.77	94.62	97.85	96.34
35	98.17	96.33	94.50	95.41	96.33	96.15
40	95.16	94.35	95.97	95.16	95.97	95.32



(a)

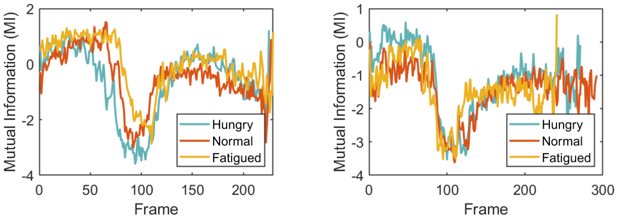
(b)

Fig. 19: Performance of SVC model: (a) four classifications; (b) binary classifications.

3) Comparison of breath-based authentication methods: We compare DBreathLock with existing breath-based authentication methods using mmWave radar (M-Auth [36]), RFID (BioTag [35]), and acoustic signals (BreathPID [14]). M-Auth and BioTag achieve non-contact breath-based authentication by capturing breath signals using a single COTS mmWave radar and two low-cost RFID tags, respectively. BreathPID achieves a contact-based method by placing the bottom of a phone on the subject's neck to capture bronchial breath sounds. In our work, DBreathLock utilizes the smartphone's acoustic modules to detect C-A-joint movements and deep breath sounds to complete authentication. As shown in Fig. 21, DBreathLock achieves slightly lower authentication accuracy than existing contact-based BreathPID, but offers performance comparable to non-contact M-Auth and BioTag, which require extra hardware and are therefore unsuitable for everyday use.

4) Performance of LDM: In this experiment, a Realme GT2 is used as the victim's device is placed on a table (case1) and held by volunteers (case2) respectively for sampling authentication data (about 80 samples) from deep breaths. We make the following assumptions: (1) these data are stolen and attackers do not need to differentiate them from case1 or case2; and (2) we use a MI 8 SE as an attack device to launch ARAs (40 samples from case1 and 40 samples from case2) at a distance of 15-20 cm from the victim device. Then, we train the SVC model based on the features of C-A-joint movements and smartphone vibrations. The train set achieves 99.11% accuracy. Moreover, Fig. 19(a) shows ROC for the test set which consists of four classes: ca1, ca2, cr1, and cr2 (discussion in IV.D). From Fig. 19(a), we can observe that AUC is 0.9444 for ca1, 0.9861 for ca2, 0.9583 for cr1, and 1 for cr2. Meanwhile, the EER for ca1 and ca2 is about 2.78%, while the EER for cr1 and cr2 is close to 0%.

In practice, ca1 and ca2 are usually considered as positive samples. cr1 and cr2 are considered as negative samples. So we also test a binary classification shown in Fig. 19(b). The AUC



(a) UserA (b) UserB

Fig. 20: Authentication capability of MI sequences.

TABLE IV: Euclidean distances of MI sequences under varying physiological states.

UAF–UAN	UAF–UAH	UAN–UAH	UBF–UBN	UBF–UBH
80	96	76	100	90
UBN–UBH	UAF–UBF	UAF–UBN	UAF–UBH	UAN-UBF
76	125	213	197	125
UAN-UBN	UAN–UBH	UAH–UBF	UAH–UBN	UAH–UBH
230	183	226	353	295

Note: UAF = UserA Fatigued, UAN = UserA Normal, UAH = UserA Hungry, UBF = UserB Fatigued, UBN = UserB Normal, UBH = UserB Hungry.

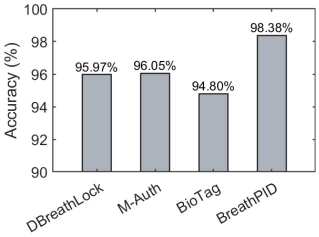


Fig. 21: Comparison of breath-based methods.

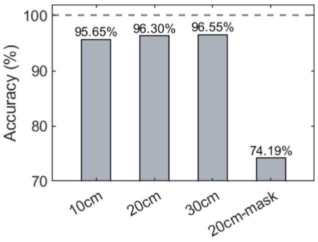


Fig. 22: Impact of detection distances and wearing masks.

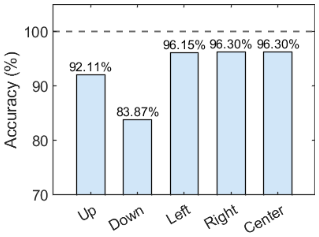


Fig. 23: Impact of smartphone orientation.

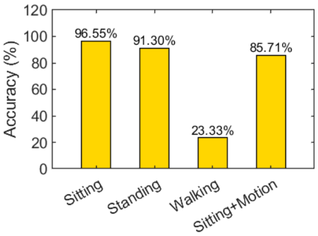


Fig. 24: Impact of different body postures.

reaches 0.9792 with only 4.17% EER. This result indicates that the LDM we designed can resist most ARAs. This is because C-A-joint movements are affected by environment-related and device-related interference during playback, and it is also difficult for attackers to fully replicate the victim's hand micro-movements.

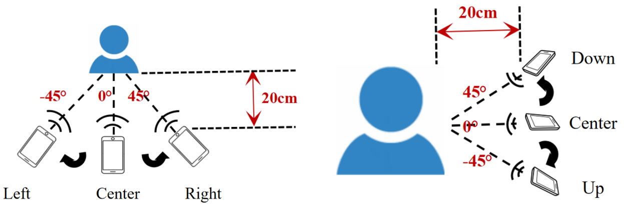
5) Authentication Capability of MI Sequences: To further explore the capability of MI sequences in authentication, we first calculate the average MI (MI_{ave}) sequence for 40 participants and select the two users whose MI_{ave} sequences appear most similar to the naked eye. $MI_{\text{ave}} = \frac{1}{n}\sum_{i=1}^{n} MI_i$, where MI_i represents the i_{th} MI sequence and n is the number of MI sequences. Then, we record the deep breath of these two users in normal, hungry, and fatigued states, and calculate the Euclidean distance of each user's MI_{ave} sequence in different states, as well as the Euclidean distance between the two users in all states. As shown in Fig. 20 and TABLE IV, the Euclidean distance between the MI_{ave} sequence of the same user in different states is significantly smaller than the distance between the MI_{ave} sequence of different users across all states. The result further indicates that the MI sequences of different users show differences, demonstrating the distinguishing capability of MI sequences in authentication.

D. Evaluation of Various Factors

This section studies the impact of various issues for practical use of DBreathLock, including wearing masks, detection distances, and orientations.

1) Impact of Detection Distances and Wearing Masks: We evaluate the impact of the detection distance between the smartphone and the volunteer who is seated. Considering the normal human length of the forearm is about 24-28 cm, we set the maximum distance to 30cm for daily usage. Fig. 22 shows the authentication accuracy of volunteers holding their smartphones at different distances ranging from 10 cm to 30 cm. The results indicate that DBreathLock's authentication accuracy all exceeds 96% at 20 cm and 30 cm. However, the distance is reduced to 10 cm with 95.65% accuracy. This counter-intuitive observation suggests that the close proximity makes the self-interference and reflected signals almost simultaneously arrive at the microphone, affecting the authentication accuracy. In some scenarios, such as hospital wards, users may wear masks. Therefore, we also test the performance of DBreathLock with volunteers wearing masks at a distance of 20 cm. As shown in Fig. 22, the authentication accuracy decreases to 74.19%. The reason is that the mask may block some of the burst sounds, thereby affecting the performance of the MSIA model.

2) Impact of Smartphone Orientation: To assess the impact of detection orientation during authentication, we adjust the angle of the smartphone relative to the chest (up, down, left, right, center), as shown in Fig. 25. The detection distance is set to 20 cm. The results, as shown in Fig. 23, illustrate that DBreathLock achieves higher authentication accuracy in the horizontal orientations (left, right, center) than in the vertical orientations (down and up). Specifically, the accuracy of DBreathLock with the down orientation decreases to 83.87%. In fact, the down orientation causes the reflected signal to contain less information about C-A-joint movements and deep breath sounds than other orientations, which may result in authentication failure. Under the up orientation, the accuracy increases to 92.11%: the bottom microphone is close to the mouth/nose and can record clearer breath sounds than in the down orientation. However, DBreathLock still cannot detect all C-A-joint movements. Therefore, to guarantee the effectiveness of DBreathLock, we employ the smartphone's built-in accelerometer to detect the pose of the smartphone and judge holding orientation. If a user holds a smartphone in a vertical orientation (up and down), DBreathLock prompts



(a) Horizontal orientations (b) Vertical orientations

Fig. 25: Different smartphone orientations.

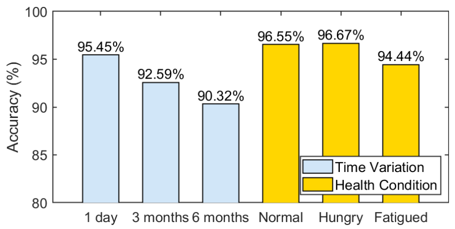
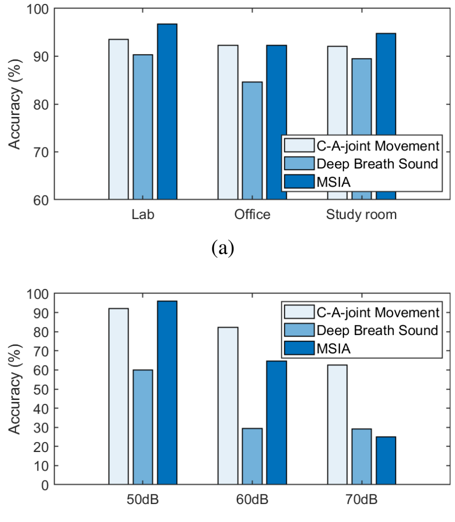


Fig. 26: Impact of different physiological states.



(a)

(b)

Fig. 27: Impact of noise: (a) under different scenarios; (b) with different decibel levels.

the user to keep the smartphone in an approximate horizontal pose for authentication.

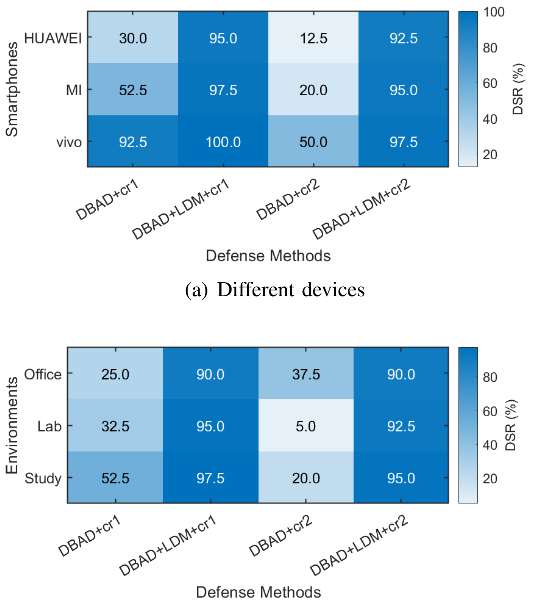
3) Impact of Different Body Postures: We study the impact of four postures on DBreathLock: sitting, standing, walking, and sitting + motion. As shown in Fig. 24, different postures do affect DBreathLock's accuracy. Specifically, when the user is sitting, the authentication accuracy reaches 96.55%. However, it drops to 91.3% when the user is standing. This is because when users keep standing, their unconscious body shaking is stronger than sitting. When the user is walking or sitting+motion, deep breath often pauses due to the body motion, causing a significant drop in accuracy. Therefore, DBreathLock uses an accelerometer to monitor smartphone movement caused by body motion and alerts the user to remain still if excessive motion is detected. Although body motion has an impact on DBreathLock, the authentication only requires the user to remain still for a short time for sampling one deep breath.

4) Impact of Different Physiological States: A person's physiological state will change over time. Therefore, it is crucial to assess the performance of DBreathLock during long-term use. We continuously monitor the breath characteristics of three users over six months. As shown in Fig. 26, the authentication performance of DBreathLock gradually decreases over time. Thus, we introduce a semi-automatic data update mechanism (active reminder + manual update) in DBreathLock to adapt to physiological changes and maintain authentication accuracy. Specifically, after a user completes authentication, the system compares the envelope similarity between the current user's deep breath and their historical deep breath data stored in the database. If the similarity falls below a predefined threshold lasting days, the system assumes that the user's physiological characteristics have changed or that the user is not breathing correctly. Even if the user passes the authentication, the system will alert the user about abnormal breathing conditions and ask them to re-record the data. If the similarity is still below the threshold, the system will confirm that the user's physiological characteristics have changed and will use the new data to fine-tune the model.

Additionally, users' health state changes also affect the performance of DBreathLock. We evaluate DBreathLock's performance under three states: normal, hungry, and fatigued. To analyze the impact of hunger and fatigue, we ask users to record deep breaths before meals and after hard work without rest, respectively. The results show that the system's performance remains almost unchanged in the hungry state; for fatigued states, the system's accuracy drops to 94.44%. It is likely that users under fatigued states may slightly reduce the amplitude of C-A-joint movements during deep breaths compared to under normal and hungry states.

5) Impact of Noise: We evaluate the performance of DBreathLock in three real scenarios: office (30-40 dB), lab (25-30 dB), and study room (20-25 dB). As shown in Fig. 27(a), the authentication accuracy of DBreathLock using C-A-joint movements in these three scenarios remains around 92-93%. However, authentication accuracy using deep breath sounds varies. In the lab and study room scenarios, the accuracy exceeds 89%, while it drops to 84.62% in the office scenario. This is mainly because deep breath sounds are relatively affected by ambient noises, whereas the features from C-A-joint movements are primarily contained in the sonar signals, making them less susceptible to noise interference.

To further investigate the impact of high-level noise on DBreathLock, we actively generate white noise at 50 dB, 60 dB, and 70 dB. Typically, the noise level of daily conversations is hard to exceed 60dB, and above 60 dB is close to busy streets. As shown in Fig.27(b), DBreathLock's authentication performance obviously decreases once the volume of noise is close to 60 dB. The reason is that high-level noise not only overwhelms breath sounds but also increases the amplitude of high-frequency harmonics which directly interfere with the sonar signal. However, solely using C-A-joint movements



(a) Different devices

(b) Different environments

Fig. 28: Impact of different devices and environments.

for authentication still achieves 82.4% accuracy. Therefore, when the system detects that the ambient noise exceeds 60 dB, it prioritizes using C-A-joint movements for authentication. Overall, DBreathLock still outperforms traditional breath sound-based authentication methods, making it suitable for daily use.

E. Evaluation of System Security

Although breaths are harder to record compared to other biometric features (such as fingerprints, faces, and voiceprints), attackers could potentially use advanced technology to hack into the victim's device and steal pre-recorded audio to complete ARAs. Besides, attackers can conduct a series of IAs by imitating deep breaths in person. Moreover, they could even combine both methods to form hybrid attacks. Therefore, in this section, we evaluate DBreathLock's performance in defending against ARAs, IAs, and hybrid attacks.

1) Impact of Advanced Replay Attacks.: As discussed in Section IV-D, different smartphone speakers have unique frequency selection characteristics, and different environments also have differences in multi-path interference. Thus, we evaluate the performance of DBreathLock in defending against ARAs in several smartphones and environments. We employ a HUAWEI nova 5i Pro, a MI 8 SE, and a vivo Y3 as attack devices, and a Realme GT2 as the victim device. Fig. 28(a) shows the DSR of DBreathLock in defending against replay attacks from three smartphones. Obviously, the performance of DBAD for three smartphones is different. For instance, the HUAWEI nova 5i Pro, MI 8 SE, and vivo Y3 have a 21.25%, 36.25%, and 71.25% chance,respectively, to successfully spoof DBAD (including cr1 and cr2). However, combined with the LDM, DBreathLock can effectively resist most replay attacks (cr1 and cr2) from all devices. Fig. 28(b) shows the DSR of DBreathLock in defending against replay attacks from three different environments (office, lab, and study room). In these three environments, DBAD can only resist up to 36.25% of replay attacks (in the study room scenario). Once combined with LDM, the DSR increases to 96.25% and 93.75% in the lab and study room scenario, respectively. In the office, because ambient noise may interfere with the sonar echoes, the DSR of DBAD+LDM drops to 90%.

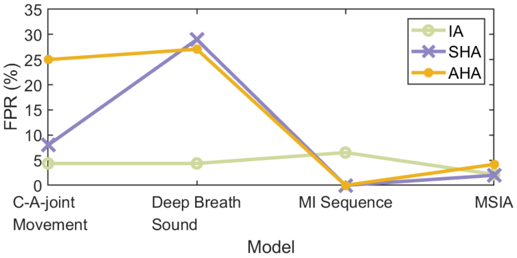


Fig. 29: Impact of IAs and hybrid attacks.

2) Impact of Impersonation Attacks: In IAs, attackers are allowed to wear headphones to listen to the legitimate user's deep breaths and launch the attacks to better grasp the characteristics of the breaths. In such attacks, fake deep breaths are from real persons, so DBAD and LDM are invalid. The resistance to such attacks mainly lies in the identity authentication model. We assess the performance of C-A-joint movement, deep breath sound, MI sequence, and MSIA models in resisting IAs. As shown in Fig. 29, DBreathLock achieves a FPR of 2.17% in defending against IAs. Actually, it is a challenge for attackers to accurately reproduce the legitimate user's deep breaths, including both the breath sounds and the related C-A-joint movements.

3) Impact of Hybrid Attacks: Fig. 29 shows the performance of DBreathLock under two kinds of hybrid attacks. C-A-joint movement model under AHAs achieves a FPR of 25% and is higher than SHAs with a FPR of 13%. The reason is that AHAs introduce the real-person C-A-joint movements to amplify the replayed sonar echoes, making it easy to fool the authentication system that only uses the C-A-joint movement model. Relying only on the deep breath sound model also results in high FPRs for both kinds of hybrid attacks, as these attacks replay the victim's deep breath sounds. However, AHAs and SHAs are difficult to fool MI sequence model. MI sequences have strict synchronization between deep breath sounds and C-A-joint movements, and it is difficult for attackers to imitate this synchronization. MSIA model achieves FPRs of 2% and 4.17% when defending SHAs and AHAs, respectively. Obviously, the C-A-joint movement and deep breath sound modalities negatively affect the final authentication in the MSIA model, but the overall performance of MSIA model is acceptable, considering the impersonation attacks.

F. System Overhead on Smartphones

We use PerfDog [62], a mobile platform performance analysis tool, to evaluate DBreathLock's performance in runtime, CPU utilization, and system power consumption of the MI

TABLE V: System overhead on different smartphones.

Phone type	Runtimes (s)	CPU (%)	Power (mAh)
MI 8 SE	1.18	19.70	14.47
Realme GT2	0.49	21.70	18.47
HUAWEI MatePad Pro	0.37	19.10	24.43

8 SE, Realme GT2, and HUAWEI MatePad Pro. Due to the constant changes in the CPU frequency of mobile devices, we adopt normalized CPU utilization that takes frequency factors into account:

$$U_{\text{norm}} = \frac{T_{\text{exec}}}{T_{\text{total}}} \times \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} f_{\max,i}}, \tag{9}$$

where U_{norm} represents the normalized CPU usage. T_{exec} is the CPU execution time. T_{total} is the total CPU runtime. f_i is the current frequency of the i_{th} CPU core. And $f_{\max,i}$ represents the maximum frequency of the i_{th} CPU core.

The experimental results are shown in TABLE V. The maximum power consumption of these three devices running DBreathLock is 24.43 mAh. Since DBreathLock is a one-time authentication system, the current battery capacity of mobile devices is sufficient to support daily operations. MI 8 SE has the longest runtime at 1.18 s, while the HUAWEI MatePad Pro has the shortest, taking only 0.37 s to complete one authentication. This demonstrates that improving the configuration of smartphones can significantly reduce runtime and even the longest runtime is still acceptable from the user's perspective. Additionally, the CPU utilization of all three devices remains around 20%, which is quite reasonable.

Overall, considering power consumption and computational resource utilization, DBreathLock can be deployed on smartphones and does not affect normal device usage.

G. Subjective Evaluation

In addition to the standard metrics evaluation, we also conduct a subjective evaluation. We invite 40 users to participate in the experiment and design a questionnaire to assess the users' acceptance of DBreathLock. The questionnaire includes the following six questions:

Q1: Are you comfortable using DBreathLock in public places (e.g., offices, labs)?

Q2: Are you comfortable using DBreathLock in quiet environments (e.g., study rooms)?

Q3: Are you comfortable using DBreathLock in multi-person environments?

Q4: How do you perceive the level of privacy protection provided by deep breath authentication compared to other biometric methods (e.g., fingerprint, face)?

Q5: Would you be willing to use deep breath authentication in various daily scenarios (e.g., login, payment, unlocking)?

Q6: Did you hear any audible sounds emitted by the device?

The questionnaire is evaluated using a Likert scale, with scores ranging from 1 (very uncomfortable) to 5 (I can accept it). A score above 3 indicates that the expected effect has been achieved. As shown in Fig. 30, the average scores for all questions are above 3. In particular, the average score for Q4 reaches as high as 4.8, indicating that most users consider deep breath authentication to offer higher privacy protection compared to other biometric methods. For Q1, Q2, and Q3, even in public or multi-person environments, the average scores remain above 3.5, showing that most users felt comfortable using deep breath authentication in these settings. Additionally, all users report not hearing any audible interference from the device during the process.

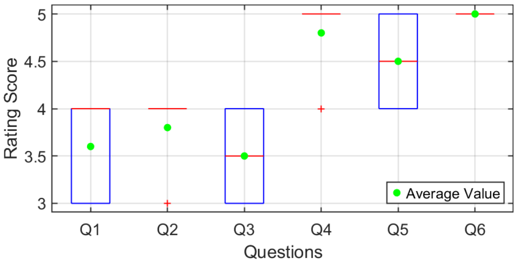


Fig. 30: Subjective evaluation after using DBreathLock.

VI. Discussion

In this section, we analyze the potential application scenarios and limitations of DBreathLock, compare it with traditional breath-based authentication systems, and discuss future directions for improvement.

1) Potential application scenarios: Different from face-based, fingerprint-based, password-based, and voice-based authentication methods for smartphones, DBreathLock is a deep breath-based non-contact authentication system for high-security scenarios. Its main advantage lies in the difficulty of recording and impersonating the personal deep breath. Specifically, DBreathLock is able to be used in online banking account access and payment services as an alternative or complementary authentication method to enhance security. Moreover, DBreathLock is suitable for scenarios where facial features are unclear (e.g., in low-light conditions) or fingerprint is damaged, which may render face-based and fingerprint-based authentication methods less effective. Furthermore, the near-silent nature of the deep breath makes DBreathLock ideal for public facilities requiring quiet environments, such as hospital wards or labs. Lastly, DBreathLock requires no hardware modifications to COTS smartphones, achieving a balance between effectiveness and practical usability.

2) Comparison with traditional breath-based authentication systems: Traditional breath-based authentication systems are vulnerable to replay and impersonation attacks. Furthermore, in noisy environments, breath sounds are easily overwhelmed by ambient noises, affecting the accuracy of traditional systems. DBreathLock, on the other hand, uses sonar signals to capture C-A-joint movements for authentication, significantly increasing the difficulty of replay and impersonation attacks. In addition, sonar signals are inaudible, so they can be emitted at maximum volume to improve the noise robustness. Finally, DBreathLock employs MI to analyze and quantify the relationship between breath sounds and C-A-joint movements. By introducing MI, our system is more adaptive to physiological changes in users (such as hunger and fatigue).

3) Limitations and Future Directions: i) Improvement in continuous authentication. Currently, DBreathLock's authentication relies on a single deep breath, which is only effective for access permission scenarios and not suitable for continuous identity authentication. To address this challenge, we plan to design an identity retention mechanism based on normal breath for continuous authentication. Actually, normal breath is more natural and difficult to observe, and thus we need to make more effort to deal with the weak signals. *ii) Impact of ambient noise.* In Section V-D, we find that once ambient noise exceeds 60 dB, the system's authentication performance decreases significantly. Therefore, in future work, we will try to employ a robust recursive least squares (RLS) [63]-based sub-band adaptive filter, which can better suppress high-frequency components of noise and thus enhance the robustness of DBreathLock.

VII. Conclusion

In this paper, we introduce a deep breath-based identity authentication system called DBreathLock. This system utilizes acoustic modules embedded in smartphones to capture the unique biometric features of the user's deep breath and verify the legitimacy of users. In this system, a smartphone emits inaudible FMCW-based sonar signals to detect C-A-joint movements and simultaneously record breath sounds. Then, a GRU-based DBAD method is developed to extract the deep breath fragment accurately. Subsequently, we present a MSIA model fused with three types of features to complete identity authentication. Additionally, to further improve the system's resistance to ARAs, we design a LDM, which utilizes features from C-A-joint movements and smartphone vibrations to differentiate between normal authentication and attacks. Extensive experiments with 40 participants in real-world environments demonstrate that DBreathLock exhibits high accuracy, robustness, and security in user authentication.

References

[1] Y. Chen, Y. Sun, B. Yang, and T. Taleb, "Joint caching and computing service placement for edge-enabled iot based on deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 501–19 514, 2022.

[2] G. Hong, B. Yang, W. Su, H. Li, Z. Huang, and T. Taleb, "Joint content update and transmission resource allocation for energy-efficient edge caching of high definition map," *IEEE Transactions on Vehicular Technology*, 2023.

[3] J. Liu, X. Zou, J. Han, F. Lin, and K. Ren, "Biodraw: Reliable multi-factor user authentication with one single finger swipe," in *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*. IEEE, 2020, pp. 1–10.

[4] A. S. Rathore, W. Zhu, A. Daiyan, C. Xu, K. Wang, F. Lin, K. Ren, and W. Xu, "Sonicprint: a generally adoptable and secure fingerprint biometrics in smart devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 121–134.

[5] K. Nguyen, H. Proença, and F. Alonso-Fernandez, "Deep learning for iris recognition: A survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–35, 2024.

[6] J. Wei, H. Huang, Y. Wang, R. He, and Z. Sun, "Towards more discriminative and robust iris recognition by learning uncertain factors," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 865–879, 2022.

[7] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: anti-spoofing facial authentication using cots rfid," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[8] R. Amadeo, "Galaxy s8 face recognition already defeated with a simple picture," *Ars Technica*, 2017.

[9] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 321–336.

[10] K. R. Alluri and A. K. Vuppala, "Iiit-h spoofing countermeasures for automatic speaker verification spoofing and countermeasures challenge 2019." in *Interspeech*, 2019, pp. 1043–1047.

[11] J. Shang, S. Chen, and J. Wu, "Defending against voice spoofing: A robust software-based liveness detection system," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2018, pp. 28–36.

[12] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*. IEEE, 2017, pp. 183–195.

[13] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 278–291.

[14] V.-T. Tran, Y.-L. Lin, and W.-H. Tsai, "Person identification using bronchial breath sounds recorded by mobile devices," *IEEE Access*, 2023.

[15] Y. Chen, M. Xue, J. Zhang, Q. Guan, Z. Wang, Q. Zhang, and W. Wang, "Chestlive: Fortifying voice-based authentication with chest motion biometric on smart devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[16] C. Pham, M.-H. Bui, V.-A. Tran, A. D. Vu, and C. Tran, "Personalized breath-based biometric authentication with wearable multimodality," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 536–543, 2022.

[17] H.-U. Jang, D. Kim, S.-M. Mun, S. Choi, and H.-K. Lee, "Deeppore: fingerprint pore extraction using deep convolutional neural networks," *IEEE signal processing Letters*, vol. 24, no. 12, pp. 1808–1812, 2017.

[18] W. Jian, Y. Zhou, and H. Liu, "Lightweight convolutional neural network based on singularity roi for fingerprint classification," *IEEE Access*, vol. 8, pp. 54 554–54 563, 2020.

[19] M. Kumar, A. Tiwari, S. Choudhary, M. Gulhane, B. Kaliraman, and R. Verma, "Enhancing fingerprint security using cnn for robust biometric authentication and spoof detection," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*. IEEE, 2023, pp. 902–907.

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[21] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, "Edgeface: Efficient face recognition model for edge devices," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[22] A. Woubie, E. Solomon, and J. Attieh, "Maintaining privacy in face recognition using federated learning method," *IEEE Access*, 2024.

[23] Y. Chen, T. Ni, W. Xu, and T. Gu, "Swipepass: Acoustic-based second-factor user authentication for smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–25, 2022.

[24] Z. Hao, Y. Wang, D. Zhang, and X. Dang, "Ultrasonicg: Highly robust gesture recognition on ultrasonic devices," in *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 2022, pp. 267–278.

[25] J. Lian, C. Du, J. Lou, L. Chen, and X. Yuan, "Echosensor: Fine-grained ultrasonic sensing for smart home intrusion detection," *ACM Transactions on Sensor Networks*, vol. 20, no. 1, pp. 1–24, 2023.

[26] W. Xu, Z. Yu, Z. Wang, B. Guo, and Q. Han, "Acousticid: gait-based human identification using acoustic signal," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–25, 2019.

[27] Y. Geng, Y. Wang, X. Wang, C. Lu, H. Yu, and M. Yuan, "Gait-based human identification using deep learning and multiple-position wearable data," in *2023 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2023, pp. 1–6.

[28] R. Chandok, V. Bhoir, and S. Chinnaswamy, "Behavioural biometric authentication using keystroke features with machine learning," in *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 2022, pp. 1–6.

[29] Y. Jiang, H. Zhu, S. Chang, and B. Li, “Mauth: Continuous user authentication based on subtle intrinsic muscular tremors,” *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1930–1941, 2023.

[30] H. Zhu, J. Hu, S. Chang, and L. Lu, “Shakein: Secure user authentication of smartphones with single-handed shakes,” *IEEE transactions on mobile computing*, vol. 16, no. 10, pp. 2901–2912, 2017.

[31] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, “Silentkey: A new authentication framework through ultrasonic-based lip reading,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–18, 2018.

[32] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, “Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals,” in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1466–1474.

[33] Z. XU, T. LIU, R. JIANG, P. HU, Z. GUO, and C. LIU, “Aface: Range-flexible anti-spoofing face authentication via smartphone acoustic sensing,” 2024.

[34] L. Wang, W. Chen, N. Jing, Z. Chang, B. Li, and W. Liu, “Acopalm: Acoustical palmprint-based noncontact identity authentication,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9122–9131, 2022.

[35] B. Hu, T. Zhao, Y. Wang, J. Cheng, R. Howard, Y. Chen, and H. Wan, “Biotag: Robust rfid-based continuous user verification using physiological features from respiration,” in *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2022, pp. 191–200.

[36] Y. Wang, T. Gu, T. H. Luan, and Y. Yu, “Your breath doesn’t lie: multi-user authentication by sensing respiration using mmwave radar,” in *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2022, pp. 64–72.

[37] J. Liu, Y. Chen, Y. Dong, Y. Wang, T. Zhao, and Y.-D. Yao, “Continuous user verification via respiratory biometrics,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1–10.

[38] S. Vhaduri, W. Cheung, and S. V. Dibbo, “Bag of on-phone anns to secure iot objects using wearable and smartphone biometrics,” *IEEE Transactions on Dependable and Secure Computing*, 2023.

[39] C. Huang, H. Chen, L. Yang, and Q. Zhang, “Breathlive: Liveness detection for heart sound authentication with deep breathing,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–25, 2018.

[40] R. Nandakumar, S. Gollakota, and N. Watson, “Contactless sleep apnea detection on smartphones,” in *Proceedings of the 13th annual international conference on mobile systems, applications, and services*, 2015, pp. 45–57.

[41] A. Bates, M. J. Ling, J. Mann, and D. K. Arvind, “Respiratory rate and flow waveform estimation from tri-axial accelerometer data,” in *2010 International Conference on Body Sensor Networks*. IEEE, 2010, pp. 144–150.

[42] S. Kraman, “The relationship between airflow and lung sound amplitude in normal subjects,” *Chest*, vol. 86, no. 2, pp. 225–229, 1984.

[43] M. Yosef, R. Langer, S. Lev, and Y. A. Glickman, “Effect of airflow rate on vibration response imaging in normal lungs,” *The open respiratory medicine journal*, vol. 3, p. 116, 2009.

[44] M. Goel, E. Saba, M. Stiber, E. Whitmire, J. Fromm, E. C. Larson, G. Borriello, and S. N. Patel, “Spirocall: Measuring lung function over a phone call,” in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5675–5685.

[45] K. Fang, J. Qiu, T. Wang, K. Zheng, L. Xing, K. Mao, and K. Chi, “Idres: Identity-based respiration monitoring system for digital twins enabled healthcare,” *IEEE Journal on Selected Areas in Communications*, 2023.

[46] T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou, “Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2959–2973, 2018.

[47] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, “Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals,” in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 54–66.

[48] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[49] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[50] Q. Chen, M. Chen, L. Lu, J. Yu, Y. Chen, Z. Wang, Z. Ba, F. Lin, and K. Ren, “Push the limit of adversarial example attack on speaker recognition in physical domain,” in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 710–724.

[51] M. Janidarmian, A. Roshan Fekr, K. Radecka, and Z. Zilic, “A comprehensive analysis on wearable acceleration sensors in human activity recognition,” *Sensors*, vol. 17, no. 3, p. 529, 2017.

[52] Z. Ali, M. S. Hossain, G. Muhammad, and A. K. Sangaiah, “An intelligent healthcare system for detection and classification to discriminate vocal fold disorders,” *Future Generation Computer Systems*, vol. 85, pp. 19–28, 2018.

[53] Y. Li, Z. Li, K. Wei, W. Xiong, J. Yu, and B. Qi, “Noise estimation for image sensor based on local entropy and median absolute deviation,” *Sensors*, vol. 19, no. 2, p. 339, 2019.

[54] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the cuidado project,” *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[55] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, “Spectral entropy based feature for robust asr,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–193.

[56] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1331–1334.

[57] N. T. T. Ho, T. B. Pedersen, L. Van Ho, and M. Vu, “Efficient search for multi-scale time delay correlations in big time series,” in *23rd International Conference on Extending Database Technology, EDBT 2020*. OpenProceedings. org, 2020, pp. 37–48.

[58] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, “Reluplex made more practical: Leaky relu,” in *2020 IEEE Symposium on Computers and communications (ISCC)*. IEEE, 2020, pp. 1–7.

[59] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[60] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. icml*, vol. 30, no. 1. Atlanta, GA, 2013, p. 3.

[61] J. Liu, C. Xiao, K. Cui, J. Han, X. Xu, and K. Ren, “Behavior privacy preserving in rf sensing,” *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 1, pp. 784–796, 2022.

[62] “Perfdog - professional mobile performance test tool,” https://perfdog.qq.com/login.

[63] T. Bahraini and A. N. Sadigh, “Proposing a robust rls based subband adaptive filtering for audio noise cancellation,” *Applied Acoustics*, vol. 216, p. 109755, 2024.

Jiefan Qiu received the B.S. and M.S. degrees from Henan Normal University, Xinxiang, and Zhejiang Normal University, Jinhua, China, in 2007 and 2010 respectively, and the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China in 2014. He is currently an associate professor in the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His current research focuses on the wireless sensing and Integrated Sensing and Communication (ISAC).

Kailu Zheng received the B.Sc. and M.S. degrees from Jilin Agricultural University, Changchun, and Zhejiang University of Technology, Hangzhou, China, in 2022 and 2025 respectively. Currently, she is working toward the PhD degree at Southeast University, Nanjing, China. Her current research interests include wireless sensing, mobile technology and digital health.

Xiyu Wang received the B.E. degree from the Zhejiang University of Technology, Hangzhou, China, in 2023. She is currently pursuing the M.S. degree with the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. Her current research interests include wireless sensing and mobile technology.

Dongfu Zhu received the B.E. degree from Inner Mongolia University of Technology, Hohhot, China, in 2021. He is currently pursuing the M.E. degree with the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His current research interests include IoT, wireless sensing and artificial intelligence.

Kaikai Chi (Senior Member, IEEE) received the B.S. and M.S. degrees from Xidian University, Xi’an, China, in 2002 and 2005, respectively, and the Ph.D. degree from Tohoku University, Sendai, Japan, in 2009. He is currently a professor in the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His current research focuses on wireless cellular network, wireless ad hoc network and wireless sensor network. He was the recipient of the Best Paper Award at the IEEE Wireless Communications and Networking Conference in 2008. He has published more than 50 referred technical papers in proceedings and journals like IEEE Transactions on Wireless Communications, IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems, etc.

Bin Yang received his Ph.D. degree in systems information science from Future University Hakodate, Japan in 2015. He was a research fellow with the School of Electrical Engineering, Aalto University, Finland, from Nov.2019 to Nov.2021. He is currently a professor with the School of Computer and Information Engineering, Chuzhou University, China. His research interests include unmanned aerial vehicle networks, cyber security, edge computing, and Internet of Things.

Tarik Taleb (Senior Member, IEEE) received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Full Professor at Ruhr University Bochum, Germany. He was a Professor with the Center of Wireless Communications, University of Oulu, Oulu, Finland. He is the founder of ICTFICIAL Oy, and the founder and the Director of the MOSA!C Lab, Espoo, Finland. From October 2014 to December 2021, he was an Associate Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. Prior to that, he was working as a Senior Researcher and a 3GPP Standards Expert with NEC Europe Ltd., Heidelberg, Germany. Before joining NEC and till March 2009, he worked as Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a lab fully funded by KDDI. From 2005 to 2006, he was a Research Fellow with the Intelligent Cosmos Research Institute, Sendai. Taleb has been directly engaged in the development and standardization of the Evolved Packet System as a member of the 3GPP System Architecture Working Group. His current research interests include AI-based network management, architectural enhancements to mobile core networks, network softwarization and slicing, mobile cloud networking, network function virtualization, software-defined networking, software-defined security, and mobile multimedia streaming.