TherapyPal: Towards a Privacy-Preserving Companion Diagnostic Tool based on Digital Symptomatic Phenotyping

Huining Li[1†], Xiaoye Qian[2†], Ruokai Ma[3], Chenhan Xu [1], Zhengxiong Li[4], Dongmei Li[5], Feng Lin[3], Ming-Chun Huang[6], Wenyao Xu[1]
[1]University at Buffalo, [2]Case Western Reserve University, [3]Zhejiang University,
[4] University of Colorado Denver, [5] University of Rochester Medical Center, [6] Duke Kunshan University
{huiningl,chenhanx,wenyaoxu}@buffalo.edu,xxq82@case.edu,{lanyouzi,flin}@zju.edu.cn
zhengxiong.li@ucdenver.edu,dongmei_li@urmc.rochester.edu,mingchun.huang@dukekunshan.edu.cn

ABSTRACT

As the demand for precision medicine rapidly grows, companion diagnostics is proposed to monitor and evaluate therapeutic effects for adjusting medicine plans in time. Although a set of clinical companion diagnostics tools (e.g., polymerase chain reaction) have been investigated, they are expensive and only accessible in a lab environment, which hinders the promotion to broader patients. In light of this situation, we take the first steps towards developing a real-world companion diagnostic tool by leveraging mobile technology. In this paper, we present *TherapyPal*, a privacy-preserving medicine effectiveness computational framework by harnessing semantic hashing-based digital symptomatic phenotyping. Specifically, sensor data captured from daily-life activities is first transformed into spectrograms. Then, we develop a hashing learning network to extract privacy-masked symptomatic phenotypes on smartphones. Afterward, symptomatic hashes at different medicine states are fed to a contrastive learning network in the cloud for treatment effectiveness detection. To evaluate the performance, we conduct a clinical study among 65 Parkinson's disease (PD) patients under dopaminergic drug treatment. The results show that *TherapyPal* can achieve around 84.1% medicine effectiveness detection accuracy among patients and above 0.925 privacy-masked scores for protecting each private attribute, which validates the reliability and security of *TherapyPal* to be used as a real-world companion diagnostics tool.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Mobile Health; Privacy-preserving; Digital Phenotyping.

ACM Reference Format:
Huining Li[1†], Xiaoye Qian[2†], Ruokai Ma[3], Chenhan Xu [1], Zhengxiong Li[4], Dongmei Li[5], Feng Lin[3], Ming-Chun Huang[6], Wenyao Xu[1]. 2023. TherapyPal: Towards a Privacy-Preserving Companion Diagnostic Tool based on Digital Symptomatic Phenotyping. In *The 29th Annual International Conference on Mobile Computing and Networking (ACM MobiCom'23), October 2–6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3570361.3592499

1 INTRODUCTION

Due to the disease heterogeneity, patients who have similar diagnoses can respond differently to the same therapeutic intervention [1]. Around 30%-70% of patients do not respond well to a particular type of medical treatment, which leads to a loss of thousands of USD medical costs for each patient [2]. To satisfy the increasing demand for precision medicine, the FDA issued in vitro companion diagnostic devices that monitor and evaluate the medicine effectiveness for adjusting treatment plans (e.g., schedule, dose, discontinuation) in time [3]. A recent report valued the global companion diagnostics market to be USD 9.9 billion by 2026 [4].

Although companion diagnostics tools have been intensively investigated in a clinical environment and have a large body of proven approaches (e.g., polymerase chain reaction [5], next generation sequencing [6], molecular imaging [7]),

[†] Both authors contributed equally to this work.

ACM MobiCom '23, October 2–6, 2023, Madrid, Spain

ACM ISBN 978-1-4503-9990-6/23/10…$15.00
https://doi.org/10.1145/3570361.3592499

they are expensive and not accessible in either daily life or primary care places, which largely impedes the promotion of companion diagnostics to a broader population.

In light of this situation, we ask: *is it possible to leverage mobile devices to extract digital symptomatic phenotype from daily-life activities (e.g., walking, talking, etc.) for developing a real-world companion diagnostic tool?* If we can, patients can be provided with always-on monitoring services of treatment effectiveness in everyday life with no burden.

This paper takes the first steps towards positively answering this question. In fact, the key to monitoring treatment effectiveness is to compare the symptom severity before and after medicine intake. A recent study has validated the feasibility of computing symptom fluctuations by monitoring daily-life activities using the smartphone's built-in sensors [8], but it is not practical to be integrated into a companion diagnostics tool for daily usage. Specifically, two key challenges remain as follows: ***1) Privacy-isolated vs. Symptomatic-preserving***: During the treatment monitoring, how to filter privacy-sensitive content from raw sensor data collected in daily-life activities, while preserving the symptom information? ***2) Diverse constitutions vs. Generic tool***: Since medicine absorption and metabolism vary from person to person, how can we develop a generic treatment effectiveness detection model that is inclusive to users with different constitutions and medical histories?

Our work unveils the opportunity of harnessing *semantic hashing in digital symptomatic phenotyping* to address the above challenges. According to the theory of semantic hashing, the hash codes obtained by a good hash function can keep the distance order in the original space as much as possible [9]. Since the medicine effectiveness detection is based on the symptom severity comparison before and after medicine, if we can preserve the symptom distance of any two sensor recordings in pairwise hashes, then the computation performed on pairwise hashes can theoretically guarantee the detection accuracy. In the meanwhile, due to the nature of hash mapping, a single hash itself is meaningless, and is impossible to reconstruct the original input, so privacy is completely protected. Additionally, one-second sensor data (with a 100 Hz sampling rate) can be compressed above 10 times with the representation of 256-bit semantic binary hashes, which is efficient to deal with daily-life longitudinal data for generic model development.

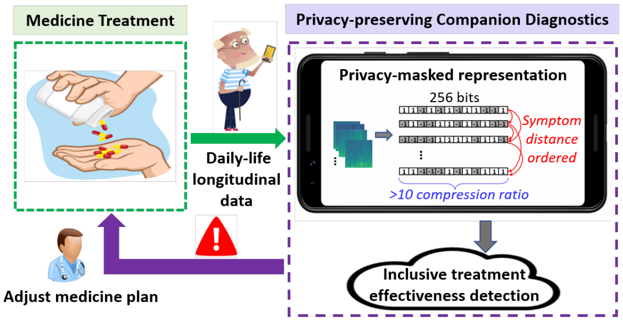

Medicine Treatment
Privacy-preserving Companion Diagnostics
Privacy-masked representation
256 bits
Symptom distance ordered
>10 compression ratio
Daily-life longitudinal data
Adjust medicine plan
Inclusive treatment effectiveness detection

Figure 1: *TherapyPal* provides patients with privacy-preserving companion diagnostic services based on privacy-masked symptomatic phenotypes.

To implement our idea in a real-world system, we present *TherapyPal*, a privacy-preserving companion diagnostic tool based on a smartphone-cloud infrastructure that monitors treatment effectiveness in daily life, as shown in Fig. 1. On the phone end, we leverage the smartphone's built-in sensors to collect activity data (e.g., gait, voice) and divide them into multiple segments. Each segment is then transformed into a spectrogram. To extract privacy-masked symptomatic phenotypes from the spectrograms of each patient locally, we first model the symptom difference by calculating the distribution divergence between feature vectors obtained from two sets of augmented spectrogram samples. Then, an extremely computation-efficient CNN-based hash learning architecture and a pair-wise loss function are designed to preserve the symptomatic structure in Hamming space. In the cloud end, triple-wise personal hashes at different medication states are fed to a contrastive learning-based hash adaptor, which transforms them into a unified symptomatic feature space that reflects the medication effect. Finally, we apply a distance threshold to the triple-wise outputs of the hash adaptor to calculate the treatment effectiveness result.

We conduct a clinical case study to evaluate *TherapyPal* on 65 Parkinson's patients under dopaminergic treatment. All the experiments are performed on unseen patients to ensure the applicability in the real world. *TherapyPal* achieves more than 82% recall and 79% precision on treatment effectiveness detection among patients with different demographic and medical backgrounds. Moreover, we perform a privacy-preserving validation study which shows the extracted hash codes can obtain an above 0.925 privacy-masked score for attributes including age, gender, and disease history.

The contribution of our work is three-fold:

- We are the first to investigate a real-world companion diagnostics tool based on activity sensing (e.g., gait, voice, screen tapping) for clinical effectiveness monitoring that satisfies the requirements of daily usage. Our exploration opens a new dimension for companion diagnostics applications and can achieve around 84.1% accuracy.
- The missing piece in mobile health is privacy-preserving analytical approaches toward daily-life longitudinal user data. We develop a generalized privacy-preserving and lightweight computational framework for sensor data representation and analysis by harnessing semantic hashing, which is not trading off privacy, analytical accuracy, and computation efficiency in mobile health applications with a theoretical guarantee.

- We design and implement *TherapyPal*, a privacy-preserving companion diagnostic tool that is inclusive to patients with diverse demographics and medical histories. We first collect sensor data and transform them into spectrograms. Then, an in-phone privacy-masked hashing learning network is applied to extract hashes that preserve the symptomatic structure. Afterward, triple-wise hashes are fed to a contrastive learning network in the cloud for predicting medicine effectiveness.

2 BACKGROUND AND PRELIMINARIES

2.1 Companion Diagnostic

Companion diagnostics leverages vitro medical devices to monitor the patient's response to treatment and examine medicine effectiveness for promoting their safety. As precision medicine rapidly develops, companion diagnostics is progressing towards the treatment of neurological disorders, cardiovascular disease, and infectious diseases [10, 11].

As the progress of the neurological disorder, nerve cells continue to deteriorate. For PD patients under chronic L-Dopa treatment, the capability of the substantia nigra cells to store dopamine would get impaired, which renders levodopa useless gradually [12, 13]. In this case, the drug effectiveness is decreased, and the symptomatic relief is impeded. Therefore, we need to harness companion diagnostics to monitor PD-related symptoms for adjusting medicine timely.

The primary goal of companion diagnostics for cardiovascular disease is to manage a patient's state (e.g., blood sugar, cardiac status) with minimum effort. For example, the glucose monitoring system is approved to obtain patients' blood sugar levels in non-clinical environments and be used to integrate with an automatic insulin dosing tool to release insulin when blood sugars exceed the normal level [14].

Abuse of antibacterial agents leads to a reduced lifespan of drugs. Companion diagnostics tests can precisely examine a patient's condition under therapeutic drugs to enable personalized adjustment of treatment plans according to their infection prognosis [15].

However, the accessibility of conventional companion diagnostics is impeded due to its intrusive diagnostics manner, expensive fee, and dependency on patients' cooperation (e.g., frequent visits to labs).

2.2 Smartphone-based Disease Symptom Measurement

With the advance of mobile technologies, many studies have investigated the diseases' symptoms via smartphone-based activity sensing. Compared with dedicated networked diagnostic medical devices or other mobile sensors (e.g., smart watches), smartphones have much higher accessibility, and the data collection and diagnosis will not create an extra burden in daily life, so they facilitate large-scale healthcare applications. For example, smartphones' built-in inertial sensors are used to collect gait data from PD patients for evaluating motor symptoms [8, 16]. Microphone-based digital biomarkers from human sounds are explored to identify respiratory symptoms and lung function [17]. Smartphones' camera-based colorimetric detection systems are developed to monitor glucose for diabetes management [18]. Inspired by these works, it is possible to leverage smartphones' built-in sensors to measure medicine-induced symptom fluctuations for realizing a companion diagnostic tool for daily usage.

2.3 Semantic Hashing Theory

Semantic hashing is widely used in image retrieval applications as it maps original input to a compact hash code for the efficiency of the nearest neighbor search [9]. In contrast to cryptographic hashing [19], the main goal of the semantic hash function is to preserve the distance order in the original input space in the hamming space. Conventional semantic hash functions leverage linear projection, kernels, spherical function, etc., to extract hash codes [20–22], but the performance is compromised for large datasets. Due to the superior representation ability of deep learning models, many studies develop deep neural networks to learn complex semantic hash functions [23, 24].

On the one side, semantic hashing maps the infinite set into finite Hamming space, so it is a theoretical "one-way" process and difficult to reconstruct the original inputs. On the other side, the learned hash code content itself is meaningless, but pairwise hash codes preserve the semantic distance of original inputs. Inspired by these, we develop a framework that will not trade off privacy and accuracy. Data are represented as semantic hashes that keep the symptom distance ordered, so the analytic performance based on symptom comparison is not degraded with a theoretical guarantee.

3 SYSTEM OVERVIEW

3.1 Application Scenario

For developers: *TherapyPal* is a general-purpose companion diagnostic platform that can provide APIs to developers. It supports diseases and symptoms that can be measured by mobile sensors, e.g., accelerometer, microphone, camera, and touchscreen sensor. For example, Alice would like to develop a real-world companion diagnostic tool for Parkinson's patients under L-Dopa treatment, but she does not know the technical details about how to extract symptomatic phenotypes for medicine effectiveness detection without leaking users' privacy. In this way, Alice can use *TherapyPal* API to configure the sensor type and model parameters. Since the *TherapyPal* API solved the privacy concerns, the developer can attract more PD patients to contribute their personal data for model training.

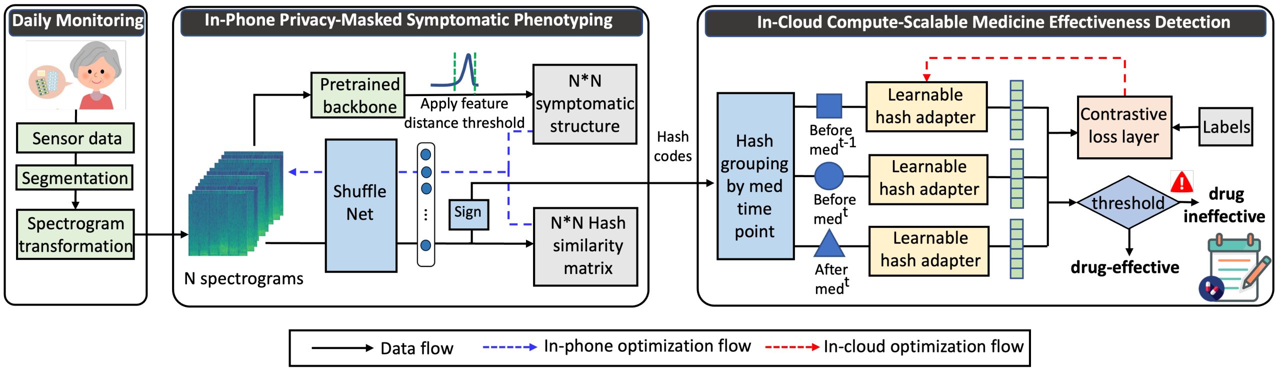

Daily Monitoring
Sensor data
Segmentation
Spectrogram transformation
In-Phone Privacy-Masked Symptomatic Phenotyping
Pretrained backbone
Apply feature distance threshold
N*N symptomatic structure
Shuffle Net
Sign
N*N Hash similarity matrix
N spectrograms
Hash codes
In-Cloud Compute-Scalable Medicine Effectiveness Detection
Hash grouping by med time point
Before med^{t-1}
Before med^{t}
After med^{t}
Learnable hash adapter
Learnable hash adapter
Learnable hash adapter
Contrastive loss layer
Labels
threshold
drug ineffective
drug-effective
Data flow
In-phone optimization flow
In-cloud optimization flow

Figure 2: The overview of *TherapyPal*, a privacy-preserving companion diagnostic tool that monitors treatment effectiveness in daily life. It leverages the smartphone's built-in sensors to collect activity data (e.g., gait, voice) and then performs privacy-masked hash learning to extract digital symptomatic phenotypes, which are finally fed to the cloud server for medicine effectiveness detection.

For end users: Bob is a PD patient under treatment for five years and has a high risk of developing drug resistance. Although the timescale of PD treatment is years, the symptoms could fluctuate on a daily basis and develop in a short time. Periodic doctor's office site visits (usually every 3-6 months per time) have two limitations, i.e., failure to select precise drugs based on patients' recall of day-to-day symptoms, and missing the optimal time to adjust medicine. In this case, he is suggested to use *TherapyPal* app and complete the activity (e.g., walking, talking) data collection before and after the medicine each day. Since *TherapyPal* masks the private attributes in activity data, it can be used as a secure tool to monitor and identify treatment effectiveness in everyday life. Once *TherapyPal* detects ineffective medicine, it will send a message to the patient's healthcare provider timely. Then, the clinical professionals will arrange an examination for patients and optimize the prescription such as drug type, daily dosage amount, and dosage time, accordingly.

3.2 *TherapyPal* Modules

As shown in Fig. 2, *TherapyPal* consists of the following three modules:

Sensor Data Collector: The data collection is in a non-clinical environment. According to the disease type, we leverage the smartphone's built-in sensors to collect the activity data, e.g., gait, voice, touchscreen tapping, and camera-based photoplethysmography (PPG). The collected sensor data are first divided into multiple segments based on data properties and structures, and then each segment is transformed into a spectrogram for preserving time-frequency characteristics in a unified 2D format. Note that one type of sensor data segment is transformed into one spectrogram. For the accelerometer, x-y-z axis data segments are connected successively and transformed into one spectrogram.

Privacy-masked Digital Symptomatic Phenotyping: The privacy-masked hash learning is performed on the obtained spectrograms of each patient locally. Without the symptom severity labels, we first model the symptom difference of the spectrograms by calculating the distribution divergence of two sets of semantics contained in a pre-trained deep architecture. Then, an extremely computation-efficient CNN-based hash learning model is developed to extract hash codes that preserve the symptomatic structure and mask the privacy attributes. Afterward, the extracted hashes (i.e., digital symptomatic phenotypes) are uploaded to the cloud server.

Medicine Effectiveness Detector: In the cloud server, hash codes at "before med^{t-1}", "before med^{t}", and "after med^{t}" (symbols are defined in Fig. 4) are grouped together and then fed to a learnable hash adaptor. By harnessing a contrastive loss function, the personalized hash codes are converted to a unified symptomatic feature space that reveals medication effects. After that, we apply a threshold on the triple-wise symptomatic feature distance to infer the medicine effectiveness. To mitigate random factors, we accumulate multiple detection results and vote to make final predictions.

4 IN-PHONE PRIVACY-MASKED SYMPTOMATIC PHENOTYPING

4.1 Challenges

We transform the sensor data into spectrograms to augment the time-frequency properties. In detail, the data is divided into multiple segments based on data properties (e.g., the gait data is segmented into gait cycles), and each segment is transformed into one spectrogram. To obtain the spectrogram, the segment is first partitioned into a series of windows with a length of 200ms for each and a 50% overlapping rate, and then the FFT operation is performed on each window.

Spectrograms contain not only behavior content but also the user's identity and health information. If we directly upload these spectrograms to the cloud for model training, privacy-sensitive information of the spectrograms may leak.

Therefore, we need to explore a secure descriptor of a spectrogram that satisfies the following requirements:

- Privacy isolated: Such a descriptor cannot be used to infer users' identity, unrelated disease, etc.
- Symptomatic-preserving: To facilitate medicine effectiveness detection, the disease symptom-related information needs to be preserved in such a descriptor.

However, these two requirements have posed a dilemma. For example, we would like our system to identify medicine effectiveness during Parkinson's treatment by understanding the pathological voice while preventing it from obtaining other disease information (such as Asthma) and biometric properties that can intrude on people's privacy. A recent work [25] proposed an adversarial training framework to impose privacy constraints during feature extraction, but it is unrealistic to label all privacy attributes.

The key to solving the aforementioned challenge is defining what we actually want our system to extract for medicine effectiveness monitoring. Given that medicine effectiveness computation exploits the medicine-induced symptom fluctuations, the symptom distance across data points should be kept in the distance across our designed descriptors, while the content in each descriptor should be unbiased in identity information. To achieve this goal, a privacy-masked symptomatic phenotyping approach based on hash learning is proposed. We first model the symptom distance structure of original data points and then design a hash learning network to preserve the symptom distance.

4.2 Symptomatic Relationship Modeling

Recent studies show that features extracted from pre-trained deep architectures have rich semantic information [26]. Therefore, we first adopt a pre-trained backbone to extract deep feature vectors from spectrograms. From a statistical view, extracted deep feature vectors for each type of semantics scatter in a space with an unknown distribution. A conventional approach that can model the semantic similarity relationship is to calculate the cosine distance for each pair of deep feature vectors, but it would introduce numerous false positives and negatives for samples located at the boundary of the distributions, which would misguide the hashing learning process. Therefore, we leverage a distribution divergence metric to measure the semantic differences of spectrograms. Specifically, we first perform random augmentations on each spectrogram, including random cropping, rotation, cutout, and Gaussian blur that do not change the spectrogram texture and frequency characteristics [27], to obtain a group of samples with the same semantics as the original spectrogram. Then, the distance between two spectrograms is estimated by calculating the sample distribution divergence of their semantics, formulated as [28]:

$$D_{jk}(\{\mathbf{u}_j^m\}_{m=1}^M, \{\mathbf{u}_k^m\}_{m=1}^M) = \frac{1}{M}\sum_{m=1}^{M}\left(\left(\frac{1}{M}\sum_{r=1}^{M}\rho(\mathbf{u}_j^m, \mathbf{u}_k^r) - \rho(\mathbf{u}_j^m, \mathbf{u}_j^r)\right)^2 + \left(\frac{1}{M}\sum_{r=1}^{M}\rho(\mathbf{u}_k^m, \mathbf{u}_k^r) - \rho(\mathbf{u}_k^m, \mathbf{u}_j^r)\right)^2\right), \quad (1)$$

where $\{\mathbf{u}_j^m\}_{m=1}^M$, $\{\mathbf{u}_k^m\}_{m=1}^M$ are the features of augmented samples of input spectrogram $\mathbf{x}_j$ and $\mathbf{x}_k$, respectively. $\rho(\mathbf{u}_j^m, \mathbf{u}_k^r)$ is the cosine distance between features.

We hypothesize that the distribution divergence among these feature vectors mainly originates from the symptom fluctuations caused by medication or other reasons in daily life, besides sensor-induced random noise. The reasons are two folds. First, the collected sensor data (e.g., inertial gait data) is not sensitive to environmental changes. Second, the feature extraction procedure is performed on the smartphone, and the obtained deep feature vectors are from the same person, so the persistent personal information (e.g., identity) is eliminated in the distance calculation. Therefore, it is feasible to assume that spectrogram pairs with distribution divergence much smaller than others have similar symptoms, and spectrogram pairs with distribution divergence much larger than others have dissimilar symptoms. To validate it, we randomly select 5 PD subjects benefiting from medicine intake. Specifically, we calculate the distribution divergence D_{ba} between two gait spectrograms collected before and after the same medication (i.e., indicating dissimilar symptoms), D_{bb} between two adjacent before-medication events (i.e., indicating similar symptoms), and D_{bo} between before-medication event and some other times (i.e., not sure whether the symptoms are similar or not). Finally, we find that $D_{ba} > D_{bo} > D_{bb}$ holds for each subject, which is consistent with our assumption.

The distribution distance (i.e., distribution divergence) histogram over all data pairs can be calculated. Then, we split the histogram from the maximum value, and approximate each part using a half Gaussian distribution [23]. The distance thresholds for symptomatic similarity and dissimilarity are calculated based on the mean and standard deviation of the Gaussian distribution in each part, respectively [23]. Based on it, we can obtain a symptomatic similarity function as:

$$S_{jk} = \begin{cases} 1 & \text{if } D_{jk} < \mu_1 - \lambda_1\sigma_1 \\ 0 & \text{if } \mu_1 - \lambda_1\sigma_1 \le D_{jk} \le \mu_2 + \lambda_2\sigma_2 \\ -1 & \text{if } D_{jk} > \mu_2 + \lambda_2\sigma_2 \end{cases}, \quad (2)$$

where μ_1 and σ_1 denote the mean and standard deviation of the Gaussian distribution for the left part and μ_2 and σ_2 for the right part. λ is a hyper-parameter to control the threshold. If the pair is symptomatically similar, the function will return 1; If the pair is symptomatically dissimilar, the function will return -1. The function is set as 0 if the ambiguous similarity is obtained.

4.3 Deep Hash Learning Network

Since binary hashing maps infinite set into finite Hamming space, it is a theoretical "one-way" process and difficult to reconstruct the original inputs. Inspired by it, we design a deep hash learning network to map spectrograms into fixed-length hash values where the binary hash itself is random and meaningless, but symptomatic relationships are preserved.

Our deep hash learning network is based on an extremely computation-efficient CNN architecture followed by a fully-collected layer with L hidden units, as shown in Fig. 3. We use the basic ShuffleNet [29] because it has two benefits: 1) the design of pointwise group convolution largely reduces computational complexity and can be deployed on mobile devices; 2) the channel shuffle operation helps the symptomatic information flow across different feature channels.

The goal of the hash learning network is to map symptomatically similar spectrogram pairs into similar hashes and map symptomatically dissimilar pairs into dissimilar hashes. We first define the hash similarity function using hamming distance, given by:

$$H_{jk} = \frac{1}{L}\mathbf{h}_j^\top \mathbf{h}_k,\ \mathbf{h}_j = sign(F(\mathbf{x}_j;\omega)), \tag{3}$$

where $F(\mathbf{x}_j;\omega)$ is L dimension output of our deep hash learning network with spectrogram $\mathbf{x}_j$ as input, ω is the learnable network parameters, $\mathbf{h}_j$ is the corresponding hash codes, and $\mathbf{h}_j \in \{-1,1\}^L$. If a pair of hash codes is similar, the hash similarity function will return a value near 1; If a pair of hash codes is dissimilar, the hash similarity function will return a value near -1. After that, we design a loss function to minimize the difference between hash similarity and symptomatic similarity of pairwise spectrograms, given by:

$$min\ \Gamma(\omega) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} CW_{jk}\cdot \|S_{jk} - \frac{1}{L}\mathbf{h}_j^\top \mathbf{h}_k\|_2^2, \tag{4}$$

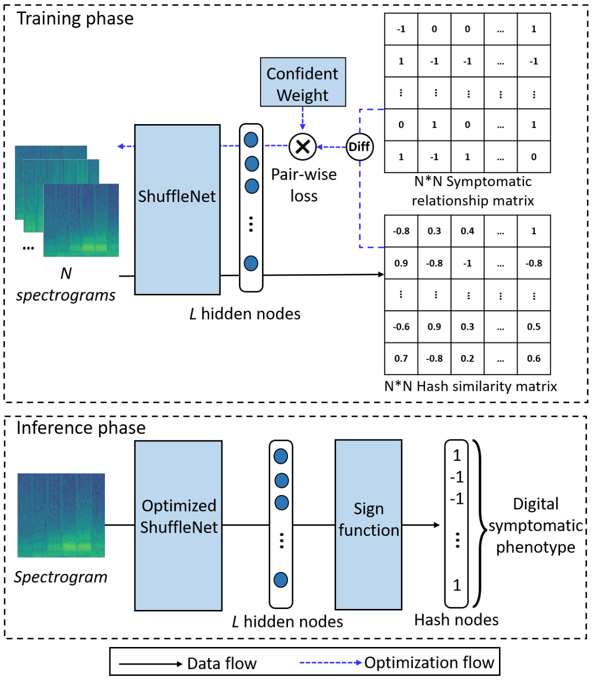
where CW_{jk} is a confident weight to guide the hash learning, which is formulated as:

$$CW_{jk} = \begin{cases} \frac{\Phi_1(\mu_1-\lambda_1\sigma_1)-\Phi_1(D_{jk})}{\Phi_1(\mu_1-\lambda_1\sigma_1)} & D_{jk} < \mu_1 - \lambda_1\sigma_l \\ 0 & \mu_1 - \lambda_1\sigma_l < D_{jk} < \mu_2 + \lambda_2\sigma_2 \\ \frac{\Phi_2(D_{jk})-\Phi_2(\mu_2+\lambda_2\sigma_2)}{1-\Phi_2(\mu_2+\lambda_2\sigma_2)} & D_{jk} > \mu_2 + \lambda_2\sigma_2 \end{cases}, \tag{5}$$

where Φ_1 and Φ_2 are the cumulative distribution function of two half Gaussian distributions, respectively, and $CM_{jk} \in [0,1]$. In our design, a potential similar pair will contribute more to the learning if D_{jk} is closer to the minimum of Φ_1, a potential dissimilar pair will contribute more to the learning if D_{jk} is closer to the maximum of Φ_2, and a pair with ambiguous similarity is not allowed for contribution.

Training. In the training phase, to enable backpropagation, we relax the binary constraint of the hashes. Specifically, the hash value is formulated as: $\widetilde{\mathbf{h}}_j = \tanh(F(\mathbf{x}_j;\omega))$. We apply the stochastic gradient descent (SGD) method [30] to minimize this function to obtain optimized network parameters.

Inference. In the inference phase, for any input spectrogram $\mathbf{x}_a$, we can generate a corresponding privacy-masked symptomatic phenotype by directly forward propagating it through the learned deep hash network as follows: $sign(F(\mathbf{x}_a;\omega))$.



-1	0	0	...	1
1	-1	-1	...	-1
⋮	⋮	⋮	⋮	⋮
0	1	0	...	1
1	-1	1	...	0

-0.8	0.3	0.4	...	1
0.9	-0.8	-1	...	-0.8
⋮	⋮	⋮	⋮	⋮
-0.6	0.9	0.3	...	0.5
0.7	-0.8	0.2	...	0.6

Figure 3: Deep hash learning network in training and inference phase.

5 IN-CLOUD MEDICINE EFFECTIVENESS DETECTOR

5.1 Problem Formulation

Preliminary. Patients benefit from medicine intake because it relieves the symptoms and slows down the disease's progress. Typically, the symptoms get relieved as a new dose of the drug starts to take effect, but they may return before the next scheduled dose [31]. This leads to a large symptom severity difference between before and after the same medication event, and a small symptom severity difference between two adjacent before-medication events. Therefore, for patients with effective medicine responses, the symptoms difference between "before med^t" and "after med^t" is more significant than the symptom difference between "before med^t" and "before med^{t-1}". For patients with ineffective medicine response, symptoms in these three states are close.

Challenge. Digital symptomatic phenotypes retain the symptom fluctuation but cannot be used directly to compute the medication effectiveness for two reasons. First, symptom fluctuations may be due to factors beyond the treatment. Hence,

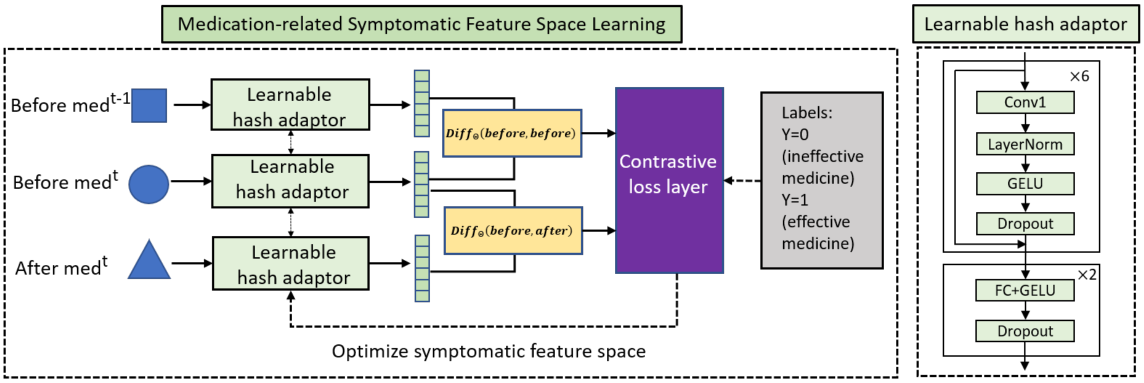

Medication-related Symptomatic Feature Space Learning
Learnable hash adaptor
Before med^{t-1}
Before med^{t}
After med^{t}
Learnable hash adaptor
Learnable hash adaptor
Learnable hash adaptor
$Diff_\Theta(before, before)$
$Diff_\Theta(before, after)$
Contrastive loss layer
Labels: Y=0 (ineffective medicine) Y=1 (effective medicine)
Optimize symptomatic feature space
×6
Conv1
LayerNorm
GELU
Dropout
×2
FC+GELU
Dropout

Figure 4: A contrastive learning-based hash adaptor that transforms personalized hash codes to a unified symptomatic feature space that reveals medication effect. “before med^t” and “after med^t” represent before and after the same medication event t; “before med^{t-1}” represents before another adjacent medication event $t-1$.

we need a supervised learning approach to specifically learn the medication-induced symptom distance. Second, since the symptomatic phenotyping is done individually, the same hash distance may represent varying degrees of symptom similarity for two patients. We need to explore a unified symptomatic feature space for model learning.

5.2 Medication-Related Feature Space Learning

Contrastive Learning-based Hash Adaptor: To solve this challenge, we design a learnable hash adaptor based on contrastive learning to transform the hashes obtained from each patient to a unified medication-related symptomatic feature space, as shown in Fig. 4. Such a hash adaptor is optimized by setting the learning task as maximizing the symptom gap between |“before med^t”–“after med^t”| and |“before med^t”–“before med^{t-1}”| for patients labeled as effective medicine response, in the same while, minimizing such gap for patients labeled as ineffective medicine response.

Specifically, the input of the hash adaptor is triple-wise hashes from the same person. We group the hashes at “before med^{t-1}”, “before med^t”, and “after med^t” together as input. The hash adaptor consists of six residual blocks followed by two MLP layers. In each residual block, the convolutional layer with 3×1 kernel size, dropout layer with drop rate 0.5, and layer norm with affine transform are applied. We employ shortcut connections in the residual block to allow the representations of different levels of processing to interact. In this way, the network can convey a high-level understanding of the last layers to the previous layers in the training phase.

To obtain a unified symptomatic feature space that can well reflect the medication effect, the contrastive loss function is designed as:

$$\begin{aligned}\mathcal{L}_{triplet} = &\frac{1}{2}(1-Y)[D_\Theta(\mathbf{h}_b^t,\mathbf{h}_a^t) - D_\Theta(\mathbf{h}_b^{t-1},\mathbf{h}_b^t)]^2 + \\ &\frac{1}{2}Y[max\{0, D_\Theta(\mathbf{h}_b^{t-1},\mathbf{h}_b^t) - D_\Theta(\mathbf{h}_b^t,\mathbf{h}_a^t) + \alpha\}]^2,\end{aligned} \tag{6}$$

Where Y is the label, $Y=0$ represents ineffective medicine response, $Y=1$ represents effective medicine response. $D_\Theta(\mathbf{h}_b^t,\mathbf{h}_a^t)$ is the learnable symptomatic distance between “before med^t” and “after med^t”, $D_\Theta(\mathbf{h}_b^{t-1},\mathbf{h}_b^t)^2$ is the learnable symptomatic distance between “before med^{t-1}” and “before med^t”, which can be formulated as:

$$\begin{aligned}D_\Theta(\mathbf{h}_b^t,\mathbf{h}_a^t) &= ||G(\mathbf{h}_b^t;\Theta) - G(\mathbf{h}_a^t;\Theta)||_2, \\ D_\Theta(\mathbf{h}_a^{t-1},\mathbf{h}_a^t) &= ||G(\mathbf{h}_a^{t-1};\Theta) - G(\mathbf{h}_a^t;\Theta)||_2,\end{aligned} \tag{7}$$

where $G(\mathbf{h}_a^t;\Theta)$ is L dimension output (i.e., unified symptomatic feature) of the hash adaptor network with hash $\mathbf{h}_a^t$ as input, Θ is the learnable network parameters. In the optimization process, if given $Y=0$, the contrastive loss function minimizes the distance between |“before med^t”–“after med^t”| and |“before med^t”–“before med^{t-1}”|; If given $Y=1$, the loss function maximizes such distance. Also, we set a margin value α that only allows the samples with effective medicine response to contribute to the contrastive loss function if such distance is within the margin.

Distance threshold-based prediction: After optimizing the hash adaptor network, we can obtain the unified medication-related symptom feature from each hash code and calculate the distance between |“before med^t”–“after med^t”| and |“before med^t”–“before med^{t-1}”|. By applying a threshold, a distance smaller than the threshold is identified as having a high risk of ineffective medicine response.

5.3 Long-term Tracking

In real practice, some random momentary events (e.g., stumbles in walking) will compromise the quality of sensor data collection. Therefore, measurement accumulation can provide a more reliable prediction. Smartphone-based sensing modality can continuously and passively track users’ activity data to obtain this goal with the minimum burden.

Since each time series measurement can be divided into multiple segments and our medicine effectiveness detection approach is performed based on segments, we further fully match these segments within the same group to generate multiple triple-wise inputs. A prediction result can be obtained from each triple-wise input. By averaging these results, we can identify whether such a group of measurements is identified as high risk of occurring drug resistance or not. After that, we accumulate the results from multiple groups and apply adaptive threshold-based voting to make the final decision on whether this patient occurring an ineffective medicine response, which can be formulated as:

$$Q = \begin{cases} \text{ineffective medicine} & \frac{1}{N}\sum_{i=1}^{N} q^i > \eta \\ \text{effectiveness medicine} & \text{Otherwise} \end{cases}, \tag{8}$$

where N is the number of groups contributed to the final decision, $q^i = 1$ if the group i is predicted as having a high risk of ineffective medicine response, otherwise $q^i = 0$, η is the voting threshold.

6 CLINICAL STUDY DESIGN

6.1 Killer Application Scenario Selection

TherapyPal is a general companion diagnostic tool that can monitor treatment effectiveness. It can be used for chronic diseases whose symptoms are measured by mobile sensors. Parkinson's disease (PD) is a common neurodegenerative disease among the elderly. One of the typical PD symptoms is difficulty walking, such as freezing gait [32], shuffling gait, and festinating gait, which can be measured by the smartphone's built-in sensors in daily activities. PD symptoms can be largely reduced after the drug intake (e.g., levodopa, dopamine agonists [33]). However, nearly 50% of PD patients develop resistance to L-Dopa treatment as the disease progresses and cells continue to deteriorate [12]. Consequently, dopaminergic drug treatment effectiveness detection among PD patients is a killer application scenario for validating the performance of *TherapyPal*.

6.2 Experimental Setup and Preparation

Participants enrollment: Our study is approved by the Institutional Review Board (IRB). We cooperate with medical centers to enroll 65 PD patients with diverse demographics who are under regular use of prescribed dopaminergic drugs in our study. These PD patients are aged from 43 to 78 years old. 36 patients are male and 29 patients are female. Their onset years are in the range of $2 \sim 21$. Under the instruction of the physicians, each subject needs to take a standard clinical drug effectiveness test [8], which is regarded as the ground truth. This test consists of questions from MDS-UPDRS [34] to comprehensively evaluate the PD patients' "off-time" experiences in daily living. The test score is scaled from 0 to 40, and a higher score indicates poorer drug effectiveness. A score less than 10 is defined as an effective drug response, and greater than 10 is defined as an ineffective drug response.

Data Collection: Data collection is conducted in a non-clinical environment. We ask subjects to install a smartphone App to collect gait data "immediately before medicine", "after medicine (when feeling best)", and "at some other time". "immediately before medicine" is 10 min before medicine. "After medicine" is one hour after medication where the frequency and amplitude of tremors are notably reduced. "some other time" is the time besides before/after medicine. For each data recording, subjects are requested to walk straight at least 20 steps while putting the smartphone in the pants' back pocket. If patients are without rear pockets, the smartphone is tied using a belt in the same position. The smartphone system provides a device-motion service to obtain the current gravity vector [35]. By contrasting the current gravity vector with the real-world gravity vector, the rotation matrix can be calculated. After multiplying the rotation matrix with measured accelerometer and gyroscope values, we can mitigate the impact of random orientations of in-pocket smartphone and obtain the data in the real-world coordinate system. Our system is performed based on the gait cycles, so each time walking recording is then segmented into multiple gait cycles. Finally, we collect 700-1000 gait cycles from each subject during the study.

Data Partition: The first step is to randomly select 50% of gait cycles for training the hash learning model for each subject. After that, the remaining gait cycles are input to the optimized hash extractor for obtaining hash codes. The second step is to train and test drug effectiveness detection model based on extracted hash codes from all subjects. To avoid overfitting, in addition to involving diverse patients, we also perform the combination and permutations on gait cycles to boost the samples. Specifically, we first group each subject's gait recordings at "before med^{t-1}", "before med^{t}", and "after med^{t}" together, and then fully match the gait cycles within the same group to generate multiple triple-wise samples. In total, we achieve 400000 triple-wise samples. Among them, three-quarters of subjects (around 300000 samples) are used for training and the remaining subjects (around 100000 samples) are used for testing. We perform four-fold cross-validation to examine *TherapyPal*'s performance.

Neural Network Implementation: The model is implemented on an NVIDIA TITAN Xp GPU by Pytorch. The mini-batch SGD optimizer is applied in a deep hashing model with 1e-5 weight decay and 0.9 momentum. As for the contrastive learning model, the Adam optimizer is utilized with 1e-5 weight decay and with a multiplicative factor of 0.9 by every 5 epochs. The learning rates of both models are 1e-3 and 1e-4, respectively. The batch size is 256.

Smartphone-end Implementation: The trained PyTorch deep hashing model is converted to an intermediate PyTorch Mobile ecosystem model in the type of PKL file and is distributed to smartphones. An HTML file is created and sent to mobile for specifying the input and parameters of the model and guides run-time execution.

Evaluation Metrics: We select recall, precision, accuracy, and F1-score as the metrics for binary classification. It is a convention to define the class of interest as positive in medical studies, and our goal is to identify ineffective risks, so we define the positive class as ineffective drug response.

6.3 Configuration

To maximize the applicability, we employ a long-term tracking module to accumulate the gait recordings for obtaining a more reliable detection result. 1/3 of patient data are used to select the optimal amount of gait recordings and the voting threshold, and the other 2/3 of patient data are for testing in the remainder of the paper. In the searching experiment, the F1-score improves and the selection of candidate voting threshold becomes broader, as the amount of gait recordings increases. The candidate range of voting threshold is nearly unchanged when the total amount reaches up to 8. With an input of 8 gait recordings, if the voting threshold is set as 3, the F1-score is more than 2.4% higher than other settings. Therefore, we choose 8 as the number of gait recordings that contribute to the medicine effectiveness detection and set the voting threshold as 3, which means the ineffective treatment result will be obtained if at least 3 of 8 triple-wise gait recordings are predicted to present ineffective drug response.

7 PERFORMANCE EVALUATION

7.1 Overall Performance

Fig. 5 shows the normalized confusion matrix of drug response detection. Overall, our system can achieve 85.1% recall, 83.4% precision, and 84.2% F1-score, which shows high diagnostic performance in digital health, especially in neurological disorder-related studies [8, 36, 37]. The recall is slightly higher than the precision, indicating that our system is better at picking up patients who do not respond well to medicine. Our system leads to 16.4% false alarm reminders that mostly occur in patients whose symptoms variations are very small between different medication states due to the lasting drug effects. If we increase the voting threshold, the false alarm rate is slightly reduced, but the missing rate largely increases, which leads to lower overall accuracy.

We further adopt a Cumulative Distribution Function (CDF) graph to demonstrate the accuracy distribution of the PD patients. Two groups of dotted lines are plotted in Fig. 6. As observed, more than 50% of PD patients achieve an accuracy above 86.4%, and only 4.7% of subjects cannot be well differentiated between good and poor medicine response. These 4.7% subjects are mostly just starting a new treatment plan or receiving DBS, so their drug effects can continue to the next medication time, we further discuss it and provide insights in Section 8.2 and Section 10. These results suggest that *TherapyPal* can achieve reliable performance on a large number of individuals.

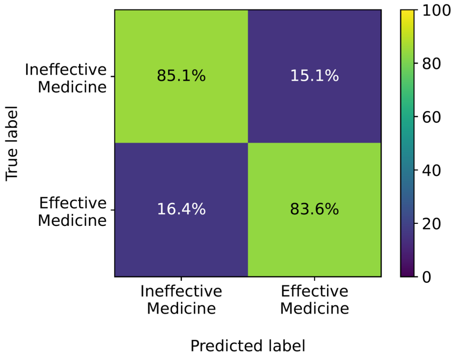


Figure 5: The normalized confusion matrix of medicine effectiveness detection.

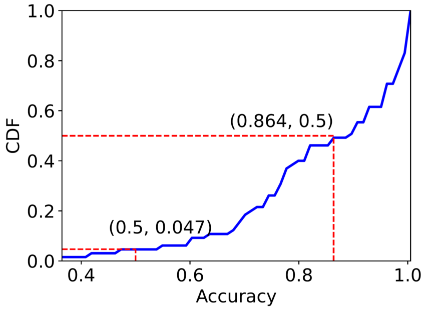


Figure 6: The CDF graph describes the medicine effectiveness detection accuracy of all PD patients.

7.2 Smartphone-end Overhead Analysis

TherapyPal is a mobile companion diagnostics tool for daily-life usage. Therefore, the smartphone-end computation should not cost many resources and can be operated with a low battery in the background process. The smartphone-end computation consists of gait cycle spectrogram transformation and a hash learning-based symptomatic phenotyping model. We evaluate the overhead on three smartphone models made in the year of 2017-2020. The experiment is set by continuously processing 20 gait cycles on the smartphone end. As shown in Table 1, the run time of a gait cycle ranges from 126 ms to 187 ms with around 40% CPU usage, which indicates that *TherapyPal* has the ability to process data in a real-time fashion. As for battery usage, the Pixel 2 consumes the most battery. If we assume that users take drugs three times a day, 10 gait cycles are collected before and after each time medicine, then Pixel 2 is estimated to use 21.6 mAh battery on our companion diagnostics tool each day, which is only 0.8% of its own 2700 mAh battery.

The Pixel 2 phone model takes 80% CPU and 1.4G memory during on-device training. In the future, we plan to employ a Tiny Training Engine [38] to reduce the training overhead.

Phone Model	Avg Runtime (ms)	Avg CPU (%)	Avg Battery Usage (mAh)
VIVO V1693A	126	40	0.31
Pixel 2	187	40	0.36
Pixel 4	165	38	0.29

Table 1: Smartphone-end overhead performance.

8 INCLUSIVENESS STUDY

8.1 Demographic Factors Analysis

Impact of Gender. Physiologic variances between males and females influence drug activity [39]. For example, females are more likely to have delayed drug effects than men. Therefore, it is critical to examine whether the medicine effectiveness detection has a bias on the gender factor. Fig. 7 shows that the gap between males and females is negligible. Specifically, the average accuracy for males and females is 84.8% and 84.1%, and the median accuracy for males and females is 89.6% and 86.5%. This is because the drug response difference induced by gender is consistent and can be mitigated by the hash adaptor. Therefore, *TherapyPal* is inclusive to the gender.

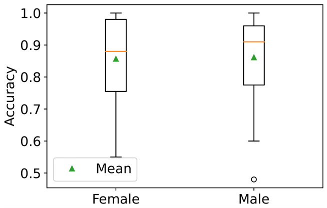


Figure 7: The influence of gender factors on *TherapyPal*'s detection accuracy.

Impact of Age. The age difference would affect the drug movement through the human body, including absorption, distribution, metabolism, and excretion [40]. As a result, we evaluate the age impacts on the medicine effectiveness detection performance. As observed in Fig. 8, the average accuracy for old-aged subjects (age > 60) and middle-aged (age ≤ 60) subjects are close, which are 85.9% and 81.8%. However, the accuracy distribution range of middle-aged subjects is wider than those of old-aged subjects. Since we have fewer data on the subjects aged below 50, the drug movement for younger patients is not sufficiently learned and causes a small number of outliers. This issue can be solved by introducing more younger subjects to train the hash adaptor. In general, *TherapyPal* is an inclusive companion diagnostics tool for patients of different ages.

8.2 Medical Factors Analysis

Impact of Treatment Duration. PD is a progressive disease that relies on dopaminergic drug treatment in a long term. Therefore, it is necessary to examine whether *TherapyPal* can be applied in different stages of the treatment. As shown in Fig 9, patients under the treatment for more than three years achieve a higher precision than those less than three years, which are 87.2% and 79%, respectively. The gap in the recall is within 1.7% between these two classes of patients. Overall, the accuracy of patients under a longer time treatment is 5.2% higher than those under a shorter time treatment. This is because the drug effect can continue to the next dosage time for the PD patients who are first initializing medication, which makes the symptom difference between before and after medicine very small and causes some false alarm cases. Since the dopaminergic drug resistance usually occurs in the middle and late stages of the treatment for PD patients, *TherapyPal* is more suggested to be used as a companion diagnostics tool starting from the middle stage.

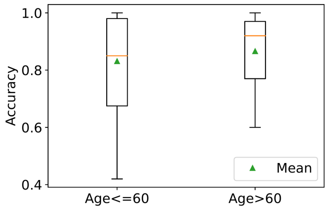


Figure 8: The influence of age factors on *TherapyPal*'s detection accuracy.

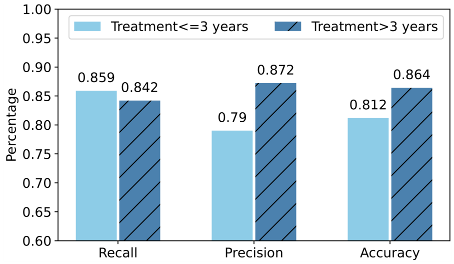


Figure 9: The influence of treatment duration on *TherapyPal*'s performance.

Impact of Disease History. Besides PD, some other diseases can also cause gait impairment. For example, concussion patients often have less single-leg stance duration, and greater stride width [41]. Stroke patients are characterized by a larger amplitude variability of steps and a relatively preserved arm swing [42]. Since our system is developed based on gait symptom fluctuation, we are curious whether non-PD-caused gait impairment will interfere with the detection. Fig. 10 shows that the difference in recall, precision, and accuracy among PD patients with different disease histories is less than 1.6 %. Such a close performance indicates that our system is robust to the complexity of gait patterns. This is because our system only captures the relative gait difference caused by medication, even though the disease histories may have unpredictable impacts on PD gait symptoms. Therefore, *TherapyPal* is a reliable companion diagnostics tool inclusive to patients with diverse health conditions.

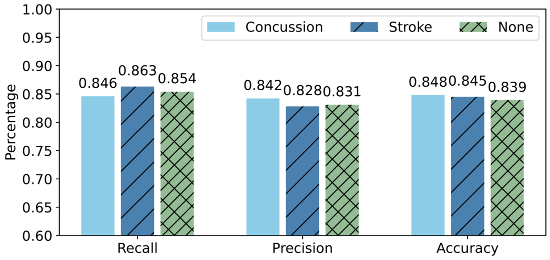


Figure 10: The influence of patients' disease history on *TherapyPal*'s performance.

Impact of Deep Brain Stimulation. Deep brain stimulation (DBS) implants electrodes within brain areas to generate the electrical impulses that regulate motor symptoms. Therefore, we are wondering whether symptom relief caused by DBS surgery will fool the medicine effectiveness detection. As observed in Fig. 11, the recall in patients with and without DBS surgery is very close, which are 85.4% and 84.3%, respectively. The precision of Patients without DBS surgery is 7.1% higher than those with DBS surgery. The reason is that our system identifies small symptom differences between before and after medicine as ineffective drug response, but the drug effect is possible to last until the next dosage time for a few patients whose medicine quality is improved by DBS surgery. Therefore, some false positive cases decrease the precision and accuracy, but the recall is not changed much. For patients with DBS surgery, we will optimize *TherapyPal* by considering the phenomenon of lasting drug effect.

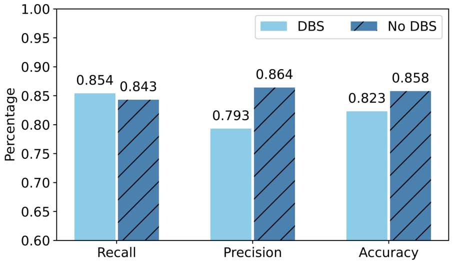


Figure 11: The influence of deep brain stimulation surgery on *TherapyPal*'s performance.

Impact of Smoking. Tobacco smoke would influence the absorption, distribution, and metabolism of the medication. Therefore, we wish to understand if such influence will compromise the detection performance. Fig. 12 shows that the recall and precision of smokers are 4.3% and 3.1% lower than those who never smoke. Our results suggest that smoking slightly affects the system performance. Since the smoking influences on medication are different from person to person, both false alarm cases and miss cases could occur, but the amount is not large. To enable a more practical companion diagnostics tool, we can request the patients to record their smoking events (e.g., timing and frequency) for developing an adaptive medicine effectiveness computation model.

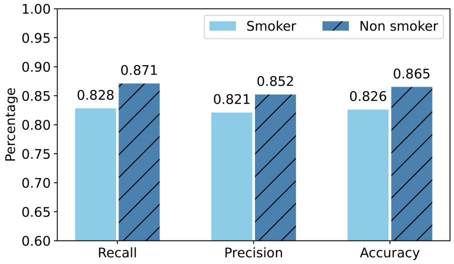


Figure 12: The influence of smoking habit on *TherapyPal*'s performance.

9 PRIVACY-PRESERVING STUDY

9.1 Setup

Data Preparation: We select three groups of PD patients. The first group is four males and four females, all in old age and having similar disease histories. The second group is four subjects with age > 60 and four subjects with age ≤ 60, who are all female and have similar disease histories. The third group is four subjects with head injury and four subjects without head injury, all female and in old age. The gait data collection and partition are the same in Section 6.2. The more variational the prepared data is, the more challenging the privacy extraction will be. Therefore, we use the data from the same medicine state (i.e., "before medicine") to create an ideal environmental condition for attackers where only privacy leakage will lead to clustering. Otherwise, the clustering could converge to before *vs.* after medicine. Totally, 20 hash samples are extracted from each subject.

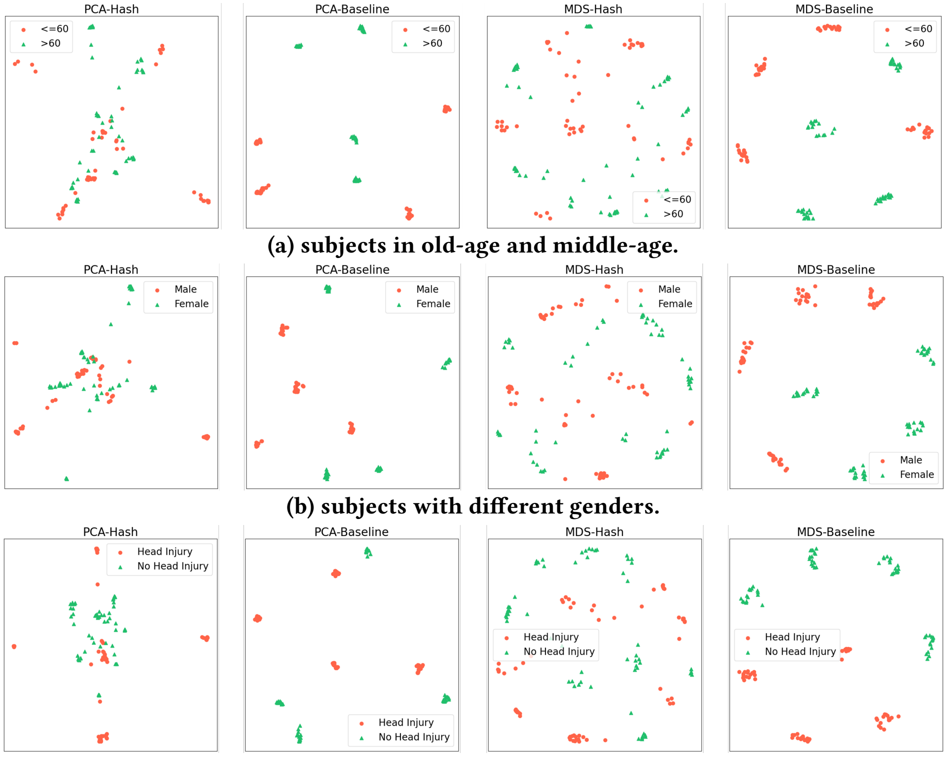
Baseline: Our baseline is a deep feature vector. Recent studies show that features extracted from deep networks are more difficult to reconstruct original inputs than those from shallow representations (e.g., SIFT) or layers. Therefore, deep feature vectors can protect users' privacy to a certain degree. In the experiment, we input each gait cycle to a ShuffleNet pre-trained on ImageNet and then extract the deep feature vector from the FC layer as our baseline sample.

Evaluation Metric: We apply feature dimensionality reduction and visualization approaches, i.e., principal component analysis (PCA) and multidimensional scaling (MDS), on hash samples and baseline samples (deep feature vectors). If the samples tend to cluster with the same identity attribute, it indicates the high risks of privacy exposure. To further quantify *TherapyPal*'s privacy-preserving capability, we introduce a Silhouette-based [43] privacy-masked score, formulated as:

$$Privacy\ Score = 1 - mean\left\{\left|\frac{b_i^{pca} - a_i^{pca}}{\{a_i^{pca}, b_i^{pca}\}_{max}}\right|, \left|\frac{b_i^{mds} - a_i^{mds}}{\{a_i^{mds}, b_i^{mds}\}_{max}}\right|\right\}, \tag{9}$$

where

$$a_i = \frac{\sum\limits_{j \in G_I, i \neq j} D(i, j)}{|G_I| - 1}, \quad b_i = \min_{J \neq I} \frac{\sum\limits_{j \in G_J} D(i, j)}{|G_J|}, \tag{10}$$



(a) subjects in old-age and middle-age.

(b) subjects with different genders.

(c) subjects with and without head injury history.

Figure 13: Visualization of feature dimensionality reduction on hash samples and baseline samples.

a_i is the mean distance between i and all other samples in the same group, b_i is the smallest mean distance of i to all samples in any other group. $\frac{b_i - a_i}{\{a_i, b_i\}_{max}}$ is the Silhouette coefficient ranging from -1 to 1. If this value is closer to 0, it means the sample is on the boundary of different identity classes, which indicates the highest uncertainty to infer user's privacy attributes. In contrast, a value closer to 1 indicates higher confidence to identify users' identity correctly, and closer to -1 indicates higher confidence to identify users' identity in a reverse manner, which is both associated with high risks of privacy exposure. Our designed privacy-masked score is calculated on the Silhouette coefficient of two PCA and MDS. It is in the range of [0, 1], and a higher score represents a higher ability of privacy protection.

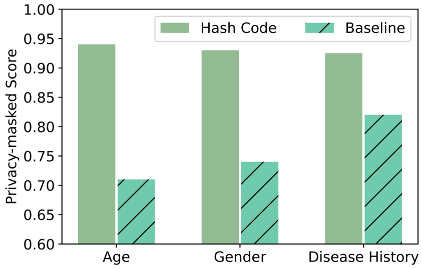


Figure 14: The comparison of the privacy-masked score for each private attribute.

9.2 Performance

As observed in Fig. 13, after applying PCA, hash samples are mixed irregularly, with some overlap between different demographics. In contrast, the baseline samples from the same subjects tend to cluster together, and each cluster is distant from the others. This observation remains consistent across all control groups. From these PCA results, we can obtain that individuals' identity attributes are prominent in baseline samples, but are well hidden in hash samples.

We also use MDS to create a visual representation of the distance between data points. Fig. 13 shows that the hash samples scatter randomly regardless of demographic similarities. Although a small number of hash samples with the same demographics are very close, they are from different subjects, resulting in a smaller distance between different subjects than between the same subject. On the other hand, baseline samples cluster by individuals, and the overall distance between different demographics is larger than that between the same demographics. Therefore, we get a consistent conclusion, i.e., hash samples pose much lower risks of privacy exposure than the baseline samples.

To further quantify system's capability of privacy protection, we calculate the privacy-masked score. As observed in Fig. 14, hash samples achieve above 0.925 privacy-masked score for each attribute, i.e., age, gender, head injury history, which is 20% higher than the baseline privacy-masked score. As the pathological properties are often implicit, the disease history has fewer exposure risks on baseline samples than other attributes. Differently, the exposure risks of all private attributes are almost the same on the hash samples, which indicates that *TherapyPal* is more neutral and generic.

10 DISCUSSION

Model Variance. Our model variance is within 4% under gender, age, and disease history factors, which is acceptable in healthcare studies. The precision variance under special conditions (e.g., just starting a new treatment plan or receiving DBS) is 7.1%-8.2% because the drug effect can last until the next dosage and cause false alarms. In the future, we plan to involve healthcare providers in the refinement of *TherapyPal* and customize the decision threshold for patients under special conditions.

Multi-Class Classification. Hashes theoretically preserve the fine-grained symptom fluctuation information, so *TherapyPal* can be extended to solve multi-class classification problems by modifying the downstream learning task in the cloud. For example, we can employ a supervised contrastive regression model [44] to learn regression-aware medication effect representation by contrasting symptomatic features against each other based on their label distances.

11 RELATED WORKS

Mobile Health Applications. Mobile health systems attracts great attention recently as they are highly accessible in daily life [45–48]. Disease screening is a typical application that leverages mobile sensors for symptom assessment. For example, SpiroSonic [49] assessed human lung function based on acoustic respiration signals. Nandakumar et al. [50] detected sleep apena events by sending frequency-modulated acoustic signals and analyzing the reflections. Another direction is medication management. PDMove [16] monitored medication adherence by analyzing medication-caused gait variability. Bae et al. [51] used a Fitbit to assess behavior risks to predict readmission for post-surgery cancer patients. PDLens [8] computed medicine-induced symptom fluctuations through daily-life activities sensing. However, among these mobile health applications, privacy protection is underexplored. Prior works (e.g., PDVocal [36]) achieved privacy-preserving through insensitive mobile data selection (e.g., breathing sounds), which limits the scope of mobile data collection. In comparison, our work is one of the first few works contributing to a new privacy-preserving analytic framework that is generalizable to broader medical applications with various data formats. In addition to the exploration of real-world companion diagnostics, we address an unmet technical need in privacy-preserving and inclusive mobile health.

Privacy-preserving Machine Learning Framework. Existing works explore many privacy-preserving machine learning models in the networked sensor systems, which mainly have two directions, i.e., *on-device federate learning* and *off-device encrypted deep learning*. In federated learning, each mobile device trains a model using personal data locally and only uploads the model parameters to the cloud for updating the global model [52, 53]. However, it is limited by upload bandwidth. Even though FedMask [54] only communicates binary masks, it needs many communication rounds for model convergence. Moreover, users' privacy can be likely recovered from uploaded model parameters [55]. A recent work [56] protects the training process inside a trusted execution environment (TEE) to avoid such recovery attacks, but the model design is constrained by the memory size of TEE. On the other side, in encrypted deep learning (e.g., CryptoNets [57], CryptoDL [58]), the mobile device encrypts the private data locally and sends it in encrypted form to the cloud, but in-cloud computation also needs to be rewritten using homomorphic operations, which limits the scalability and efficiency of machine learning models. In contrast, our privacy-preserving analytic framework can theoretically guarantee privacy as well as computation accuracy and efficiency toward daily-life longitudinal user data.

12 CONCLUSION

In this paper, we propose a privacy-preserving mobile companion diagnostics tool, *TherapyPal*, to monitor the treatment effectiveness through daily-life activities sensing. *TherapyPal* is based on a privacy-preserving computational framework that leverages semantic hashing for privacy-masked symptomatic phenotyping. Specifically, smartphone's built-in sensors collect behavior data and transform them into spectrograms. Then, we model the symptomatic relationship based on distribution divergence between extracted deep feature vectors from spectrograms. A hashing learning network is further designed to mask the privacy attributes while preserving symptom distance. In the cloud end, with triple-wise hashes (at different medicine states) as input, we develop a contrastive learning-based hash adaptor to map personalized hashes to a unified medication-related symptomatic feature space. Finally, we apply a threshold on the symptomatic feature distance for medicine effectiveness detection. Our experiments show that *TherapyPal* can achieve above 80% accuracy on medicine effectiveness detection among patients with different backgrounds and medical histories. A privacy-preserving validation study is further performed to examine the security of *TherapyPal* for daily usage.

ACKNOWLEDGMENTS

We thank all anonymous reviewers for their insightful comments on this paper.

REFERENCES

[1] J. T. Jørgensen and M. Hersom, “Companion diagnostics—a tool to improve pharmacotherapy,” *Annals of Translational Medicine*, vol. 4, no. 24, 2016.

[2] V. Wurcel, O. Perche, D. Lesteven, D.-A. Williams, B. Schäfer, C. Hopley, R. Jungwirth, A. Postulka, R. Pasmans, L.-L. Hermansson, *et al.*, “The value of companion diagnostics: overcoming access barriers to transform personalised health care into an affordable reality in europe,” *Public Health Genomics*, vol. 19, no. 3, pp. 137–143, 2016.

[3] C. for Devices and R. Health, “Companion diagnostics.” [Online]. Available: https://www.fda.gov/medical-devices/in-vitro-diagnostics/companion-diagnostics

[4] “Companion diagnostics market.” [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/companion-diagnostics-market-155571681.html?gclid=CjwKCAjw7e_0BRB7EiwAlH-goGo5sJnfpc2mORkPEx9WbM0-CLdoeBIRLsC2fN3Cv3SySm1k_SvaOhoC6CkQAvD_BwE

[5] J. L. Ramirez and J. R. Kratz, “Quantitative polymerase chain reaction for companion diagnostics and precision medicine application,” in *Companion Diagnostics (CDx) in Precision Medicine.* Jenny Stanford Publishing, 2019, pp. 55–71.

[6] S. Pant, R. Weiner, and M. J. Marton, “Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics,” *Frontiers in oncology*, vol. 4, p. 78, 2014.

[7] A. D. Puranik, H. R. Kulkarni, and R. P. Baum, “Companion diagnostics and molecular imaging,” *The Cancer Journal*, vol. 21, no. 3, pp. 213–217, 2015.

[8] H. Zhang, G. Guo, C. Song, C. Xu, K. Cheung, J. Alexis, H. Li, D. Li, K. Wang, and W. Xu, “Pdlens: smartphone knows drug effectiveness among parkinson's via daily-life activity fusion,” in *Proceedings of the 26th annual international conference on mobile computing and networking*, 2020, pp. 1–14.

[9] X. Luo, H. Wang, D. Wu, C. Chen, M. Deng, J. Huang, and X.-S. Hua, “A survey on deep hashing methods,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2020.

[10] V. Valla, S. Alzabin, A. Koukoura, A. Lewis, A. A. Nielsen, and E. Vassiliadis, “Companion diagnostics: State of the art and new regulations,” *Biomarker Insights*, vol. 16, p. 11772719211047763, 2021.

[11] M. R. Trusheim and E. R. Berndt, “The clinical benefits, ethics, and economics of stratified medicine and companion diagnostics,” *Drug Discovery Today*, vol. 20, no. 12, pp. 1439–1450, 2015.

[12] J. Nonnekes, M. H. Timmer, N. M. de Vries, O. Rascol, R. C. Helmich, and B. R. Bloem, “Unmasking levodopa resistance in parkinson's disease,” *Movement Disorders*, vol. 31, no. 11, pp. 1602–1609, 2016.

[13] B. Thanvi and T. Lo, “Long term motor complications of levodopa: clinical features, mechanisms, and management strategies,” *Postgraduate medical journal*, vol. 80, no. 946, pp. 452–458, 2004.

[14] D. Huckle, “The impact of new trends in pocts for companion diagnostics, non-invasive testing and molecular diagnostics,” *Expert Review of Molecular Diagnostics*, vol. 15, no. 6, pp. 815–827, 2015.

[15] P. J. Dailey, T. Elbeik, and M. Holodniy, “Companion and complementary diagnostics for infectious diseases,” *Expert Review of Molecular Diagnostics*, vol. 20, no. 6, pp. 619–636, 2020.

[16] H. Zhang, C. Xu, H. Li, A. S. Rathore, C. Song, Z. Yan, D. Li, F. Lin, K. Wang, and W. Xu, “Pdmove: Towards passive medication adherence monitoring of parkinson's disease using smartphone-based gait assessment,” *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 3, no. 3, pp. 1–23, 2019.

[17] X. Xu, E. Nemati, K. Vatanparvar, V. Nathan, T. Ahmed, M. M. Rahman, D. McCaffrey, J. Kuang, and J. A. Gao, “Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–22, 2021.

[18] Ö. Kap, V. Kılıç, J. G. Hardy, and N. Horzum, “Smartphone-based colorimetric detection systems for glucose monitoring in the diagnosis and management of diabetes,” *Analyst*, vol. 146, no. 9, pp. 2784–2806, 2021.

[19] R. Sobti and G. Geetha, “Cryptographic hash functions: a review,” *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 2, p. 461, 2012.

[20] J. He, W. Liu, and S.-F. Chang, “Scalable similarity search with optimized kernel hashing,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 1129–1138.

[21] A. Gionis, P. Indyk, R. Motwani, *et al.*, “Similarity search in high dimensions via hashing,” in *Vldb*, vol. 99, no. 6, 1999, pp. 518–529.

[22] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, “Spherical hashing,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2012, pp. 2957–2964.

[23] E. Yang, C. Deng, T. Liu, W. Liu, and D. Tao, “Semantic structure-based unsupervised deep hashing,” in *Proceedings of the 27th international joint conference on artificial intelligence*, 2018, pp. 1064–1070.

[24] E. Yang, T. Liu, C. Deng, W. Liu, and D. Tao, “Distillhash: Unsupervised deep hashing by distilling data pairs,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2946–2955.

[25] Z. Wu, Z. Wang, Z. Wang, and H. Jin, “Towards privacy-preserving visual recognition via adversarial training: A pilot study,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 606–624.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[28] X. Luo, D. Wu, Z. Ma, C. Chen, M. Deng, J. Huang, and X.-S. Hua, “A statistical approach to mining semantic similarity for deep unsupervised hashing,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4306–4314.

[29] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[30] M. Li, T. Zhang, Y. Chen, and A. J. Smola, “Efficient mini-batch training for stochastic optimization,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 661–670.

[31] C. D. Marsden and J. D. Parkes, “" on-off" effects in patients with parkinson's disease on chronic levodopa therapy,” *The Lancet*, vol. 307, no. 7954, pp. 292–296, 1976.

[32] R. Iansek, F. Huxham, and J. McGinley, “The sequence effect and gait festination in parkinson disease: contributors to freezing of gait?” *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, no. 9, pp. 1419–1424, 2006.

[33] P. M. C. Group *et al.*, “Long-term effectiveness of dopamine agonists and monoamine oxidase b inhibitors compared with levodopa as initial

treatment for parkinson's disease (pd med): a large, open-label, pragmatic randomised trial," *The Lancet*, vol. 384, no. 9949, pp. 1196–1205, 2014.

[34] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[35] Apple, "Getting processed device-motion data." [Online]. Available: https://developer.apple.com/documentation/coremotion/getting_processed_device-motion_data

[36] H. Zhang, C. Song, A. Wang, C. Xu, D. Li, and W. Xu, "Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds," in *The 25th annual international conference on mobile computing and networking*, 2019, pp. 1–16.

[37] Y. Gao, Y. Long, Y. Guan, A. Basu, J. Baggaley, and T. Ploetz, "Towards reliable, automated general movement assessment for perinatal stroke screening in infants using wearable accelerometers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–22, 2019.

[38] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, C. Gan, *et al.*, "On-device training under 256kb memory," in *Advances in Neural Information Processing Systems*.

[39] H. P. Whitley and W. Lindsey, "Sex-based differences in drug activity," *American family physician*, vol. 80, no. 11, pp. 1254–1258, 2009.

[40] A. A. Mangoni and S. H. Jackson, "Age-related changes in pharmacokinetics and pharmacodynamics: basic principles and practical applications," *British journal of clinical pharmacology*, vol. 57, no. 1, pp. 6–14, 2004.

[41] D. N. Martini, M. J. Sabin, S. A. DePesa, E. W. Leal, T. N. Negrete, J. J. Sosnoff, and S. P. Broglio, "The chronic effects of concussion on gait," *Archives of physical medicine and rehabilitation*, vol. 92, no. 4, pp. 585–589, 2011.

[42] K. F. de Laat, A. G. van Norden, R. A. Gons, L. J. van Oudheusden, I. W. van Uden, B. R. Bloem, M. P. Zwiers, and F.-E. de Leeuw, "Gait in elderly with cerebral small vessel disease," *Stroke*, vol. 41, no. 8, pp. 1652–1658, 2010.

[43] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[44] K. Zha, P. Cao, Y. Yang, and D. Katabi, "Supervised contrastive regression," *arXiv preprint arXiv:2210.01189*, 2022.

[45] T. Hao, G. Xing, and G. Zhou, "isleep: Unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013, pp. 1–14.

[46] N. Bui, A. Nguyen, P. Nguyen, H. Truong, A. Ashok, T. Dinh, R. Deterding, and T. Vu, "Pho2: Smartphone based blood oxygen level measurement systems using near-ir and red wave-guided light," in *Proceedings of the 15th ACM conference on embedded network sensor systems*, 2017, pp. 1–14.

[47] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 351–360.

[48] H. Zhang, G. Guo, E. Comstock, B. Chen, X. Chen, C. Song, J. Ajay, J. Langan, S. Bhattacharjya, L. A. Cavuoto, *et al.*, "Rehabphone: a software-defined tool using 3d printing and smartphones for personalized home-based rehabilitation," in *Proceedings of the 18th international conference on mobile systems, applications, and services*, 2020, pp. 434–447.

[49] X. Song, B. Yang, G. Yang, R. Chen, E. Forno, W. Chen, and W. Gao, "Spirosonic: monitoring human lung function via acoustic sensing on commodity smartphones," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[50] 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.

[51] S. Bae, A. K. Dey, and C. A. Low, "Using passively collected sedentary behavior to predict hospital readmission," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 616–621.

[52] L. Tu, X. Ouyang, J. Zhou, Y. He, and G. Xing, "Feddl: Federated learning via dynamic layer sharing for human activity recognition," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 15–28.

[53] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing, "Clusterfl: a similarity-aware federated learning system for human activity recognition," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 54–66.

[54] A. Li, J. Sun, X. Zeng, M. Zhang, H. Li, and Y. Chen, "Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 42–55.

[55] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 691–706.

[56] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "Ppfl: privacy-preserving federated learning with trusted execution environments," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 94–108.

[57] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *International conference on machine learning*. PMLR, 2016, pp. 201–210.

[58] E. Hesamifard, H. Takabi, and M. Ghasemi, "Cryptodl: Deep neural networks over encrypted data," *arXiv preprint arXiv:1711.05189*, 2017.