FEATURE ARTICLE

Smartphone Health Biomarkers: Positive Unlabeled Learning of In-the-Wild Contexts

Abdulaziz Alajaji, Walter Gerych, Luke Buquicchio, Kavin Chandrasekaran, Hamid Mansoor, Emmanuel Agu, and Elke A. Rundensteiner, *Data Science, Worcester Polytechnic Institute, Worcester, MA, 01609, USA*

There has recently been increased interest in context-aware mobile sensing applications due to the ubiquity of sensor-rich smartphones. Our DARPA-funded Warfighter Analytics for Smartphone Healthcare (WASH) project is exploring passive assessment methods using smartphone biomarkers and context-specific tests. Our envisioned context-specific assessments require accurate recognition of specific smartphone user contexts. Existing context datasets were either* scripted *or* in-the-wild*. Scripted datasets have accurate context labels but user behaviors are not realistic. In-the-wild datasets have realistic user behaviors but often have wrong or missing labels. We introduce a novel coincident data gathering study design in which data were gathered for the same contexts using both a scripted and in-the-wild study. We then propose positive unlabeled context learning (PUCL), a transductive method to transfer knowledge from highly accurate labels of the scripted dataset to the less accurate in-the-wild dataset. Our PUCL approach for context recognition outperforms state-of-the-art methods with an increase of over 3% in balanced accuracy.

Warfighters face an increased exposure to various ailments from traumatic brain injury (TBI) to infectious diseases such as COVID-19, which can adversely impact their long-term health and missions. Despite this increased exposure, warfighters are currently assessed either by healthcare providers or with intrusive and unreliable self-reports. Both are administered too infrequently, are inadequate, and can have result in severely negative outcomes. For instance, there were nearly 414 000 reports of TBI among service members between 2000 and 2019[1] and veterans with a history of TBI are 1.55 times more likely to commit suicide.[2]

Given the scale of the problem, the Defense Advanced Research Projects Agency (DARPA) has invested in research to create smartphone-based methods to assess TBI and infectious diseases early.[3] Over 81% of US adults now own smartphone.[4] Their rich sensors, processing power, programmability, and ubiquitous ownership make smartphones viable platforms for passive, continuous assessment of the health condition of its owner. Sensors built into smartphones include accelerometers, gyroscopes, light meters, and GPS. Data from these sensors can be analyzed using machine learning for health assessment. Prior work has demonstrated that smartphones can be utilized to asses a variety of ailments including depression,[5] influenza,[6] cardiovascular disease, fall detection, and heart failure monitoring.[7]

Passive health assessments often leverage sensed digital *smartphone biomarkers*, which are smartphone-sensable user behaviors that reliably indicate the health status and symptoms of the smartphone user. Smartphone health biomarkers leverage ubiquitous computing methods drawn from areas such as human activity recognition (HAR) and human context recognition (HCR).[8] Smartphone health biomarkers capture anomalous patterns in the user's behaviors, physiological signs, activities, and context visit patterns. For example, in comparison to a healthy user, a user with influenza might become more sedentary, spend more time lying in bed, and cough more. Capturing smartphone health biomarkers relies on

1536-1268 © 2021 IEEE
Digital Object Identifier 10.1109/MPRV.2021.3051869
Date of publication 12 February 2021; date of current version 9 March 2021.

accurate and reliable HCR, which this article focuses on. The actual health assessments will be the focus of future work.

Context definition: We define human context as a tuple $<$Activity, Prioception, App Usage, Social$>$. *Activity* refers to the user's current activity (e.g., walking, running). *Prioception* is the phone placement or position on its user (e.g., in front pocket, bag), which can affect the sensor readings.[9] *App usage* is the category of app, the user is currently using (e.g., communications app, social media), which is predictive of human health-relevant behaviors such sedentary levels.[10] *Social* refers to how many people the user is currently with (e.g., alone, in public).

The problem: Existing HCR and HAR datasets have limitations that make them unsuitable for our work. *1) Many existing HCR datasets are not large enough:* for training robust, generalizable machine learning HCR models. This is due to the high interperson variance in human behavior, and is especially true for deep learning, which requires millions of records for training. *2) Datasets have either inaccurate labels or unrealistic user behaviors:* HCR datasets are collected using study designs that are either *scripted*[11] or *in-the-wild*.[9] *Scripted* studies are typically conducted in laboratory setting. Participants perform scripted tasks in a fixed, predetermined order while a smartphone app continuously records reading from smartphone sensors. Human proctors annotate the users' data with corresponding context labels. In *unscripted* ("or in the wild") studies, data are collected for days in the real world as subjects live their lives. A smartphone continuously records smartphone sensor data continuously as subjects live their lives. Periodically, subjects annotate their data with labels of the contexts they visited. While the scripted method for HCR data collection yields extremely accurate and consistent labels suitable for the supervised machine or deep learning, the contexts visited and sensor data collected in each context are not representative of real life. In-the-wild HCR studies yield more realistic data. However, some of the context labels may be missing as users forget to label when they get busy with their lives. Some labels may also be wrong due to human labeling errors.[12]

Thus, from a machine learning perspective, the context labels in datasets generated by unscripted, in-the-wild studies are weak and biased toward the specific contexts the user visited. For instance, desk workers will have more instances of "sitting at desk" labels than a construction worker. Bias also occurs because certain contexts are easier to label than others. For instance, "sitting at desk" is easier to label than "swimming," a hands-free activity. The labeling quality also depends on how conscientious the study subject is, which is variable. Such poor, biased, and varied quality of context labels pose a significant challenge for machine and deep learning algorithms.

Prior work to mitigate the poor labeling quality: To mitigate missing labels, some prior smartphone data collection methods include interfaces and mechanisms that subjects can utilize to label batches of past contexts retrospectively whenever they have free time.[9] However, as subjects often do not remember some contexts or their start/end times accurately, recall bias diminishes the quality of labels. Finally, careless subjects may also provide many wrong labels.[13] Zeni *et al.*[14] devised an interactive machine learning framework for testing user trustworthiness by checking the consistency of the user provided annotations using available ground truth. Their work required continuous feedback from the user, which is undesirable and focused on user location only.

Our approach for improving the context labeling quality: We propose methods to mitigate the above-mentioned challenges, overcome the poor label quality, and improve deep learning HCR models. We propose a novel coincident data collection protocol in which data are gathered data for the same contexts of interest using both a scripted and in-the-wild design. We also propose positive unlabeled context learning (PUCL) a deep learning framework that leverages this coincident context datasets. The coincident study tries to combine the accuracy of the scripted labels with the realistic context visit patterns of the in-the-wild studies. PUCL improves HCR model performance on the realistic but noisy in-the-wild data using intelligence learned from the high fidelity labeling in the scripted dataset. Specifically, PUCL uses a transductive positive unlabeled (PU) learning methodology to transfer knowledge from the highly accurate labels of the scripted dataset to the less accurate, more sparsely but yet more realistic in-the-wild dataset. Our methodology that combines coincident data collection with PUCL outperforms state-of-the-art deep learning HCR methods and is inspired by meta-learning approaches.[15]

The scope of this article: We focus on recognizing specific contexts in which high-specificity TBI and infectious diseases assessments can be performed on monitored smartphone users. We refer the reader to the background section for a more detailed description of our envisioned health application's use case. Details about the actual TBI and Infectious disease

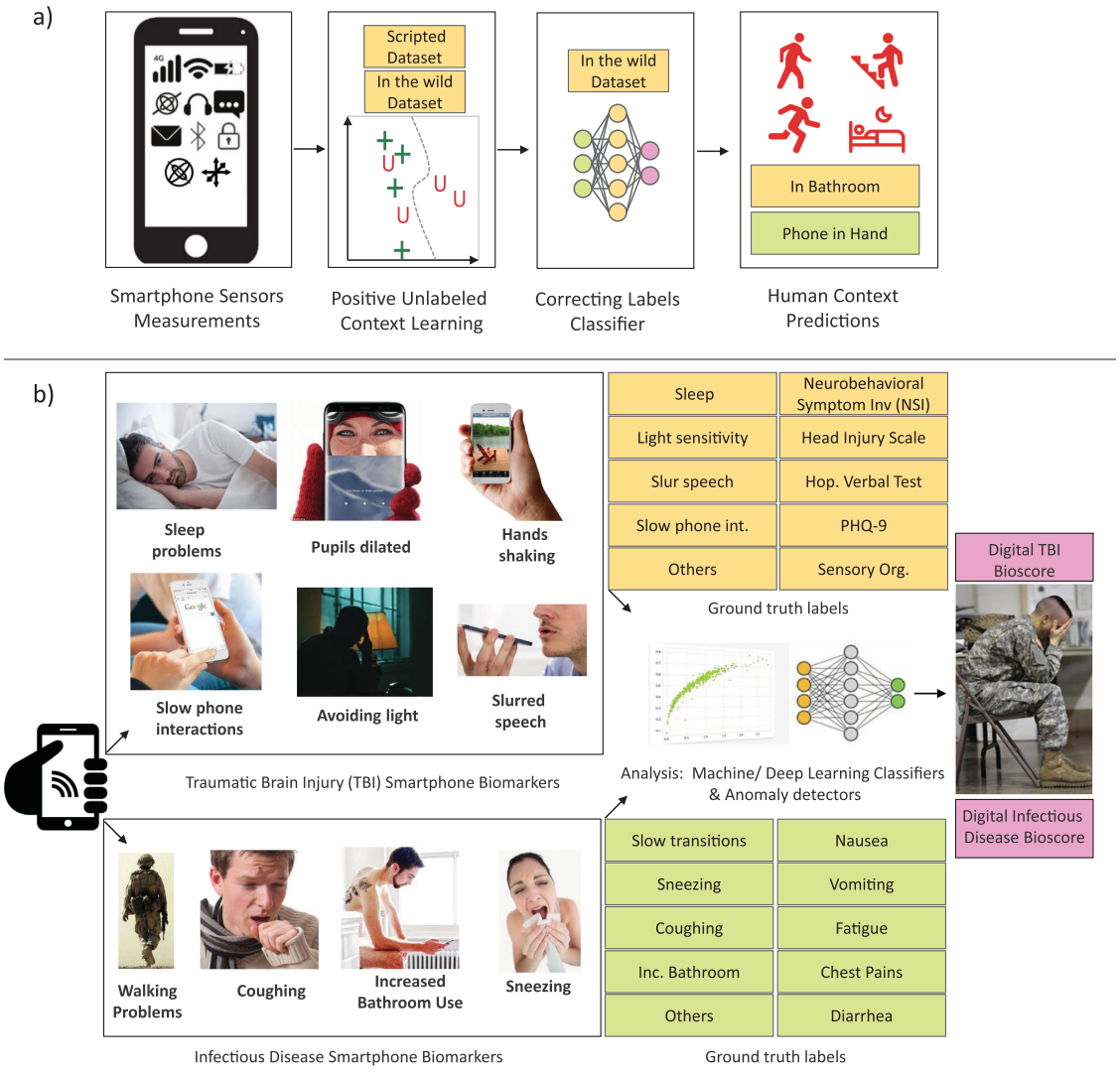

FIGURE 1. (a) High-level overview of this work, showing the scope of this article. (b) Our envisioned future health application, an overview of WASH-WPI's TBI & Infectious Disease BioScore Synthesis.

assessments are also described in the background section. For example, shaking hands (tremors) is a common symptom of TBI and other diseases.[16] If the Warfighter Analytics for Smartphone Healthcare (WASH) sensing application accurately recognizes a user's context to be $< *, \text{Phone In Hand}, *, * >$ (with "*" denoting a wildcard), then a context-specific test could assess whether their hand is shaking (tremors) by analyzing data from their phone's accelerometer and gyroscope sensors. For infectious diseases such as COVID, increased usage of the bathroom, coughing, and sneezing are all symptoms that the WASH app will assess. This article focuses only on context recognition. Research into the actual context-specific ailment assessments is not covered. But will be explored in future work. In Figure 1, we show a high-level overview of our work, in addition to the envisioned healthcare application's use case.

RELATED WORK

HAR and HCR Work

Our definition of context includes the user's activity and, thus, relates to research in HAR, a well-researched topic.[R1–R8] As more data have become available, deep-learning-based HAR methods have become popular.[R1–R5] However, the majority of these prior approaches assume that the user is only performing a single activity while the phone is carried or placed in one position on the user (e.g., in front

pocket). Such deep learning models only predict a single activity per data instance. In essence, they perform *multiclass* classification of each instance.[R9]

However, in real life, users often perform multiple activities concurrently (e.g., talking while walking) and their phones can be placed in a variety of pockets. More recent work such as *ExtraSensory*[R10] and *DeepSense* [R4] allow each data instance to be assigned with *multiple* labels and the smartphone to be carried in several possible locations (or bags/pockets). Such work frame the HAR or HCR problem as a *Multilabel* classification problem.

Context Recognition Data Gathering Studies

Several smartphone-labeled HAR and context datasets have been collected for creating machine and deep learning models and publicly released.[R9,R11–R13] The majority of these datasets were collected using study designs that were mostly scripted and typically gathered data using one or a few smartphone models that proctors used for the study. Analyzing data from diverse smartphones is important as prior work has found that sensor readings for the same activity or context can vary by up to 30% across smartphone models. In such cases, models trained on data from only one smartphone model does not generalize well to other smartphones.[R15]

Consequently, more recent datasets such as the *ExtraSensory dataset*[9] were gathered using a more realistic *in-the-wild* study design. Participants installed data collection apps, which collected data passively from their smartphones and smartwatches simultaneously as they lived their lives. Periodically, subjects were prompted to annotate activities and contexts they visited using the *ExtraSensory app*. In addition to being more realistic, these in-the-wild studies gathered data using subjects' smartphones that reflected diverse manufacturer hardware. While the *ExtraSensory* dataset came close to meeting our project's needs, several labels were sparse. Out of 100 labels defined by the investigators, subjects provided labels for only 51 contexts in 2 weeks of participation time. Moreover, this dataset did not provide labels for all the contexts we aimed to recognize for our infectious disease and TBI tests.

PU Learning

In PU learning, only a subset of the dataset is labeled. The training dataset has some labeled positive examples $\mathcal{P}$ and a set of unlabeled examples $\mathcal{U}$ that is a mixture of positive and negative examples. The goal is to create a binary classifier that can classify previously unlabeled instances or correct wrong labels in a test set into the positive or negative class.[R16] PU Bagging[R16] is an ensemble method that creates a set of trees. Each tree is trained on a subset of the entire training set and is used to classify the positive instances from the unlabeled instances. Next, each tree scores all the remaining instances in the in-the-wild dataset except the instances it was trained on. The average prediction probability from all trees for each in-the-wild instance is then given as the probability that the instance is of the positive class.

Knowledge Transfer for Labeling Sensor Data

Several semisupervised[R17,37] and transfer learning approaches[R19,R20] have previously been proposed to tackle the issue of limited annotated sensor data. Chen *et al.*[37] utilized ensemble learning and majority voting for semisupervised learning, using a similar feature generation mechanism to ours that uses attention to focus on salient regions in sensory inputs. Recently, an opportunistic sensor data knowledge transfer labeling mechanism was proposed, which leverages computer vision mechanism to label sensor-based instances. However, it requires the availability of activity recorded using a camera.[R20] Additionally, the generative auto-encoder-based method has been utilized for stochastic feature generation, utilized for cross-sensor classification of wearable data.[R19] However, these prior work were applied on HAR datasets, containing only singly labeled scripted activity data.[R19] Our work focuses instead on phone context, which includes activity but also includes other variables such as the phone's placement. To the best of our knowledge, ours is the first work to apply PU learning on coincident scripted and in-the-wild smartphone datasets to improve HCR performance. Our work demonstrates that data collected in laboratory settings can be used to improve performance of classifiers designed to infer context from data gathered in the wild. Our methods do not require the use of external devices such as cameras for annotation purposes, or interactive correction of wrong labels by humans.

BACKGROUND

DARPA WASH Program

The DARPA-funed Warfighter Analytics using Smartphone for the Healthcare (WASH) DARPA project[17] is investigating passive smartphone assessment of TBI and infectious disease. This will provide an up-to-date assessment of the warfighter's battle readiness.

TABLE 1. A) Summary of target contexts. we focused only on recognizing the labels corresponding to the "physical state" and "phone prioception" subcategories of the context tuple defined early. for our experiments, we sampled data with these labels from both the scripted and in-the-wild datasets. b) context-specific ailment tests to detect tbi and infectious diseases and relevant human contexts.

a)

Target Contexts			
Laying Down, Phone on Table	Exercising, Phone in Pocket	Toilet, Phone in Pocket	Walking, Phone in Bag
Walking, Phone in Hand	Walking, Phone in Pocket	Typing	Sleeping
Sitting	Running	Laying Down (state)	Standing
Talking On Phone	Bathroom	Phone in Pocket	Phone in Hand
Phone in Bag	Phone on Table, Facing Up	Phone on Table, Facing Down	Stairs - Going Up
Stairs - Going Down	Walking		

b)

Traumatic Brain Injury	
Ailment Test	**Test Context**
Worse Reaction Time	< Interacting with Phone, in Hand, *, * >
Increased Light Sensitivity	< *, in Hand, *, * >
Unilateral Pupil Dilation	< Interacting w/ Phone, in Hand, Texting, * >
	< Interacting w/ Phone, in Hand, Video chat, * >
Hands Shaking	< *, in Hand, *, * >
Slurred Speech	< Talking into Phone, *, *, * >
Infectious Diseases	
Ailment Test	**Test Context**
Increased Cough Frequency	< Coughing, *, *, * >
Increased Sneezing	< Sneezing,*, *, * >
Resting Heart Rate	< Sitting, in Pocket, *, * >
Increased Toilet use Frequency	< Using Toilet, *, *, * >
Change in respiration	< Sleeping, on Table, *, * >
	< Exercising, *, *, * >
Both TBI and Infectious Disease	
Ailment Test	**Test Context**
Increase In Activity Transition Time	< Lying down, Phone In Pocket, *, * >
	< Sitting, Phone In Pocket, *, * >
	< Standing, Phone In Pocket, *, * >
Change in Sleep Quality	< Sleeping, *, *, * >
Change in Gait	< Walking, Phone in Pocket/Hand, *, * >

Target populations include active duty service members and veterans initially but discoveries made will also apply to civilians.

In the envisioned use case, the WASH smartphone app will passively gather smartphone sensor data throughout each day. Each day of data is then pushed to the cloud overnight for analyses. In the cloud, disease inference models will analyze this data to generate a *bioscore* (or probability of illness) for each warfighter.

Program phases: The WASH program is divided into two distinct phases. Phase 1 involves recognizing specific smartphone user contexts in which targeted health assessments will be conducted. Phase 2 involves creating the methods for the actual TBI and infectious disease assessments of smartphone users. In phase 1, we researched and created a list of smartphone biomarkers that were predictive of TBI and infectious diseases and corresponding contexts. Our team conducted user studies to collect labeled data for those contexts and created HCR models to infer those contexts from labeled smartphone sensor data. Table 1(a) shows the list of contexts such as "walking, phone in hand" that our team gathered labeled data on and created HCR models. The planned ailment-specific tests or biomarkers corresponding to each of these contexts is listed in Table 1(b). We created our list of ailment tests and contexts in consultation with

TBI and infectious disease experts from the University of Massachusetts Medical School (UMMS). As a concrete example, shaking hands is a sign of TBI. In phase one, our team conducted user studies and created deep learning models to detect that the smartphone user was holding their phone. In phase 2, we are focusing on assessing whether the user's hand is shaking.

DARPA WASH PHASE 1: OUR CONTEXT RECOGNITION APPROACH

Novel Coincident Context Data Gathering Study

In our novel coincident study design approach, we conducted both scripted and in-the-wild gathering studies to gather labeled data in the same contexts as described in Table 1(a). Our in-the-wild context study was similar to the *Extrasensory* study approach. The smartphone app continuously gathered sensor data on 103 subjects' phones as they lived their lives. Users were then prompted to self-report labels of contexts they visited. Our scripted study was conducted in a specific laboratory, buildings, or routes on our campus. The smartphone app collected data from 100 participants that visited the listed contexts in Table 1(a) in a scripted fashion. The scripted data-gathering session lasted approximately 1 h per subject, and human proctors oversaw and manually annotated the data. Both datasets (scripted and in-the-wild) were preprocessed and featurized using the same approach detailed in the paper by Alajaji *et al.*[18] Segments were generated using a window size of 10 s and contexts were handled as multi-label vectors. Each context was assigned a binary label (e.g., walking versus not walking). For our proposed methods to work, it is required that both datasets have the same number of binary classes.

Background: Types of Supervised Learning

In supervised learning tasks such as classification and regression, predictive models are trained on annotated training examples. A training example is comprised of an input feature vector (or instance) and an associated label (or ground-truth). Weak (or inaccurate) labels can occur, requiring innovative learning methods. There are three categories of weakly supervised learning: 1) *Incomplete supervision:* utilizing unlabeled training data; 2) *Inexact supervision:* only coarse-grained labels are provided; and *3) Inaccurate supervision:* where the labels are not always true.[19] In many practical scenarios such as in our in-the-wild HCR study, it is challenging to gather an adequate amount of high-quality labels for fully supervised learning due to the high costs of gathering labeled data. Consequently, learning methods that can be trained under weak supervision are desirable.[19] In this work, we try to improve the accuracy of recognizing the inexact and weakly supervised in-the-wild context data using a PU learning approach to learn the correct labels of data instances that have wrong or missing labels from labels in the scripted (supervised) data. We use a regularization technique inspired by meta-learning approaches. Meta-learning is a subfield of machine learning focusing on learning from prior experience.[15] Specifically, we are inspired by the dual meta-learner & learner techniques, leveraging meta knowledge from a supporting set with highly accurate labeled dataset. This guides the learning process on the main learning task on the weakly labeled dataset.[15] Our approach is similar to the paper by Dehghani *et al.*[15] in leveraging prior knowledge from a highly accurate labeled dataset.

Challenges: Two major challenges arise, which must be addressed. In our in-the-wild dataset, the labels are extremely noisy, inaccurate, and sparse. Second, it is challenging to devise a robust methodology for transferring knowledge from the scripted dataset to the much noisier, yet more true-to-life, in-the-wild dataset. Specifically, discovering the most likely true labels of mislabeled or unlabeled scripted data is a very challenging problem.

Positive Unlabeled (PU) Context Learning (PUCL): A Novel Learning Methodology

To tackle the abovementioned challenges, we propose *Positive Unlabeled (PU) Context Learning (PUCL)*, a novel learning methodology that has two stages and is depicted in Figure 2. In the first stage, we utilize a PU classifier with our correctly labeled scripted dataset to correct inexact or coarse labels in our in-the-wild context dataset. In the second stage, we train DeepContext, our novel deep learning architecture, on the in-the-wild dataset with labels that have been corrected by our PU method during the first stage. We now describe the two stages of PUCL in more detail.

Stage 1: Correcting the In-the-Wild Labels

In this stage, PUCL tries to learn reliable label-feature mappings from the more reliable scripted dataset, allowing us to discover incorrect or missing labels in

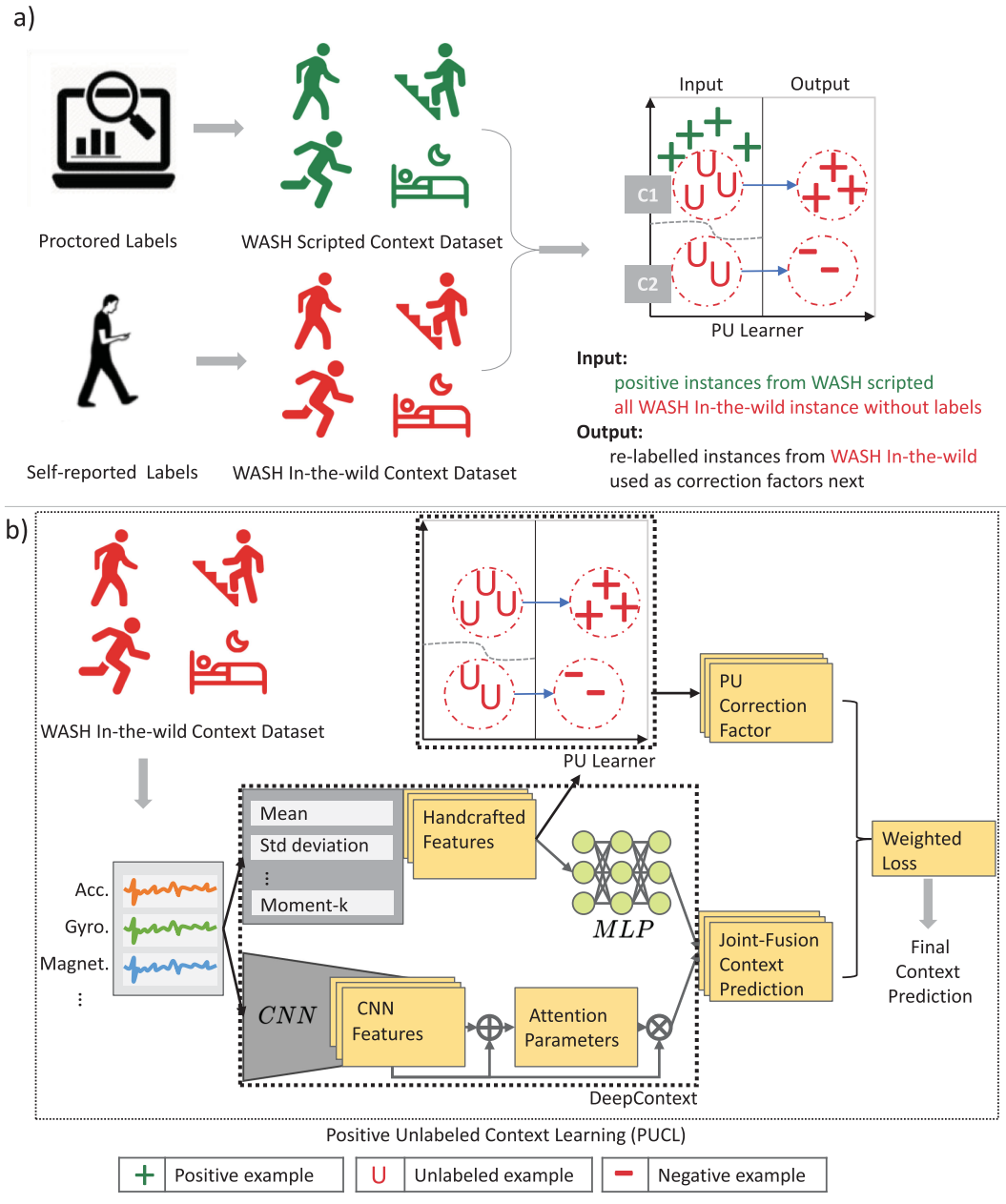

FIGURE 2. Diagram for our positive unlabeled context learning (PUCL) showing a) PU learning for coincident scripted and in-the-wild HCR. A PU learner is fit to identify (+) and (−) instances in the WASH In-the-wild dataset. C1 instances are relabeled as positives due to their proximity to positively labeled instances of WASH Scripted Context Dataset in feature space2). C2 instances are relabeled as negatives due to their distance from positive instances of the WASH Scripted Context Dataset. b) DeepContext HCR model is trained using the pseudolabels generated as a correcting factor.

the in-the-wild dataset. For each class Y that is present in both the scripted and in-the-wild dataset, let $\mathcal{P}$ be the positively labeled instances of that class in the scripted dataset. Let $\mathcal{U}$ be the entirety of the in-the-wild dataset. We then train a probabilistic PU classifier f_{pu} to predict $Pr(Y = 1|x \in \mathcal{U})$.

While our approach is flexible enough to utilize any PU learning method, we use the PU Bagging algorithm as our classifier. After running the PU Bagging algorithm, all in-the-wild instances would have now been associated with a probability of belonging to the positive class. In addition to guessing the labels of

unlabeled instances, our PUCL method can also correct wrongly labeled instances in the in-the-wild dataset. Positive instances that are wrongly labeled as negative can be identified because the PU bagging algorithm will assign them a score that indicates that they have a high probability of belonging to the positive class. Conversely, negative instances that are wrongly labeled as positive will have a score assigned by the PU bagging algorithm, which indicates that they have a low probability of belonging to the positive class.

Building on this intuition, we formulate and estimate a *correcting factor* that corresponds to how much an assigned label in the in-the-wild dataset should be trusted. For each class, e.g., "'walking," let y_i be the label of that class associated with the ith instance in the in-the-wild dataset and let PU_i be the PU Bagging score for the class associated with that instance. Then, the correcting factor CF_i is given as follows:

$$\mathrm{CF}_i = 1 - |y_i - \mathrm{PU}_i|.$$

If PU_i is large while $y_i = 0$, or if PU_i is small while $y_i = 1$, then CF_i will be close to 0. This means that the label associated with the ith instance should not be trusted.

Stage 2: Context Recognition Using DeepContext

The goal of this stage is to train a robust context classifier, *DeepContext*,[18] a novel deep learning based architecture for multilabel recognition of a smartphone user's current context. Utilizing an attention mechanism, *DeepContext* is able to autonomously learn salient features that discriminate contexts, while suppressing potentially irrelevant parts of the input.

We adapt *DeepContext*, our proposed context classification model but additionally we utilize PUCL to mitigate the negative impact of the inaccurate and missing labels in the in-the-wild dataset. Specifically, *DeepContext* uses the correcting factor in stage 1 to improve its classification results on the in-the-wild dataset. *DeepContext* takes as input both handcrafted-features generated using domain knowledge as well as the raw-sensor values collected by the smartphone. Furthermore, DeepContext utilizes state-of-the-art attention mechanisms to focus on subcomponents of the input data that are most predictive of each target class.

Classification accuracy is boosted as the noise present in each input is ignored. In particular, DeepContext is trained using gradient descent on its parameters (denoted as Θ) on the inexact and weakly labeled in-the-wild data by minimizing the following cost function:

$$C(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_\Psi(f(X_i), Y_i)$$

where N is the number of training samples and ℓ_Ψ is the loss function that is weighted by the correcting-factor. More specifically,

$$\ell_\Psi = \sum_{y\in Y}\sum_{i=1}^{N}\left(\frac{1}{Pr(y)} + \mathrm{CF}_{i,y}\right) \cdot [y_i \log(f(x_i) + (1-y_i)\log(i - f(x_i))].$$

More intuitively, ℓ_Ψ is simply the cross entropy loss (or any other deep learning loss function) multiplied by a weighting factor. The weighting factor is a combination of the inverse class frequency with the correcting factor. In order to account for class imbalance, the weighting factor weights instances of infrequent classes higher than instances of frequent classes and discounts the cost incurred from instances that are likely to have been mislabeled by the annotator. This discounting is applied so as not to punish the network for disagreeing with annotator-assigned labels that are probably wrong.

Context Recognition Results

We compared our model performance against state of the art for HCR (ExtraSensory MLP[9]), which has been applied for very similar dataset to ours. Due to class imbalance in our context datasets, we utilize *Balanced Accuracy* (BA) as the metric to evaluate the context recognition performance of *DeepContext* and our novel PUCL method. BA is defined as follows:

$$\mathrm{BA}(\mathcal{D}) = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)$$

which is also

$$\mathrm{BA}(\mathcal{D}) = \frac{1}{2}(\mathrm{Sensitivity} + \mathrm{Specificity}).$$

Our results described in Table 2(a) demonstrate that our approach outperforms the state-of-the-art model on all metrics except recall with lower false positive rates. We speculate that the drop in recall may be due to the amount of mislabeled annotated true positives. Intuitively, the PU Correcting Factor will put less attention on instances that are most likely mislabeled. It would then be expected that the true positive instances will be classified with a higher consistency using guided knowledge gathered from the scripted dataset. Table 2

TABLE 2. Comparison of our results with state-of-the-art methods.

a) Results overall					
Model	BA	Recall	Precision	Specificity	F1-score
ExtraSensory MLP	0.780,161	**0.781,946**	0.339,242	0.778,377	0.472,944
PU Context Learning (PUCL)	**0.813,777**	0.713,059	**0.551,373**	**0.914,494**	**0.621,843**
b) Results per label					
Label	ExtraSensory	*PU Context Learning*			
Phone on Table, Facing Down	0.8416 ± 0.014	**0.8707 ± 0.011**			
Stairs - Going Down	0.8051 ± 0.012	**0.8429 ± 0.011**			
Sleeping	0.8294 ± 0.002	**0.8419 ± 0.005**			
Stairs - Going Up	0.8141 ± 0.002	**0.8414 ± 0.005**			
Laying Down, Phone on Table	0.7913 ± 0.018	**0.8204 ± 0.010**			
Phone in Bag	0.7924 ± 0.018	**0.8024 ± 0.010**			
Phone in Pocket	0.7878 ± 0.014	**0.7994 ± 0.016**			
Typing	0.7736 ± 0.033	**0.7945 ± 0.008**			
Walking, Phone in Bag	0.7763 ± 0.010	**0.7910 ± 0.007**			
Walking, Phone in Pocket	0.7740 ± 0.008	**0.7895 ± 0.010**			
Phone one Table, Facing Up	0.7563 ± 0.011	**0.7888 ± 0.015**			
Walking	0.7602 ± 0.003	**0.7796 ± 0.003**			
Exercising, Phone in Pocket	0.7532 ± 0.025	**0.7711 ± 0.012**			
Laying Down	0.7410 ± 0.021	**0.7701 ± 0.009**			
Walking, Phone in Hand	0.7519 ± 0.006	**0.7654 ± 0.010**			
Sitting	0.7408 ± 0.008	**0.7577 ± 0.018**			
Running	**0.7551 ± 0.058**	0.7539 ± 0.029			
Phone in Hand	0.7436 ± 0.009	**0.7512 ± 0.016**			
Bathroom, Phone in Pocket	0.7246 ± 0.019	**0.7476 ± 0.010**			
Jogging	**0.7720 ± 0.145**	0.7437 ± 0.175			
Exercising	0.7185 ± 0.037	**0.7429 ± 0.008**			
Jumping	0.7176 ± 0.004	**0.7260 ± 0.002**			
Standing	0.7037 ± 0.010	**0.7256 ± 0.002**			
Bathroom	0.6614 ± 0.024	**0.6957 ± 0.004**			

(b) presents performance per label, showing that our approach consistently outperforms state-of-the-art methods across all labels, except Running and Jogging. As it can be seen in the results, Running and Jogging have the highest variability scores among user splits. The poor performance on Running and Jogging can be resulted from the increased noise resulted from such activities. Figure 3(a) and (b), we compare confusion matrices showing that our approach achieves consistent improvements over other state-of-the-art methods in detecting phone prioception (placement). Last, in Figure 3(c), we evaluated the impact of the proposed PUCL mechanism on both DeepContext and ExtraSensory MLP. We show that utilizing the PU Correcting factor during training, we achieved a significant increase in the BA of classification in our evaluation on both learning models: ExtraSensory MLP and our proposed model DeepContext.

PROJECT OUTLOOK

The Warfighter Analytics using Smartphone for Health (WASH) DARPA project is exploring smartphone sensing methods to passively assess the TBI and infectious disease status of active duty service members and veterans. Findings in the program will likely be useful for also assessing civilian populations. In this work, we introduced a novel PUCL approach for applying transductive PU learning on coincident scripted and in-the-wild HCR datasets. In future work, we plan to exploit unsupervised joint feature and label representation methods to further improve the accuracy on this challenging task. Moreover, one of the additional steps, we could do to improve our PUCL approach is to consider the amount of the introduced bias due to the different data distributions in both scripted and in the wild datasets. We plan to explore the limitations of our PUCL approach to such coincident data gathering study design. Finally, we are also researching the passive biomarker and health tests for TBI and infectious diseases.

ACKNOWLEDGMENTS

This work was supported in part by the Computer Science Department at Worcester Polytechnic

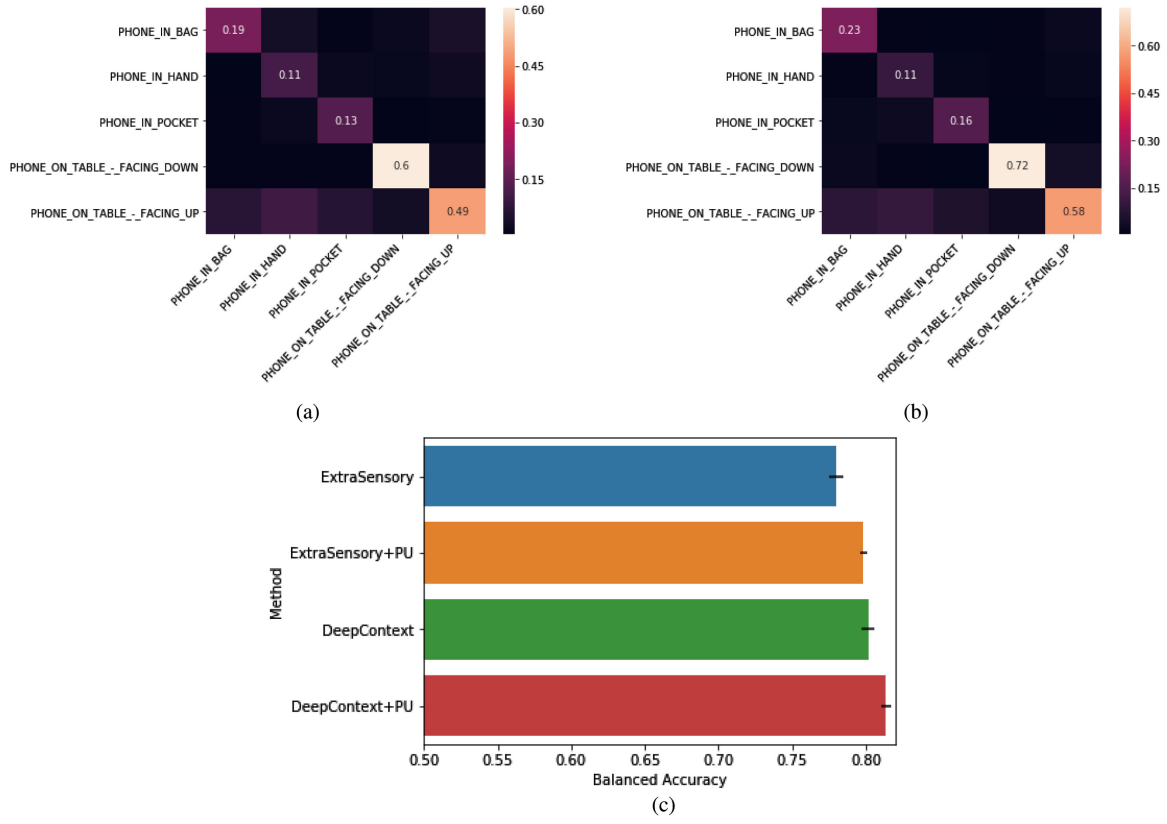

FIGURE 3. (a) and (b) Confusion matrices for phone prioception, with normalized scores. (c) Impact of PU correcting factor on the used learning method.

Institute and the DARPA WASH under Grant HR00111780032-WASH-FP-031and in part by DARPA under Agreement FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

1. DVBIC. Dod Worldwide Numbers for TBI. Accessed: Jun. 7, 2020. [Online]. Available: https://dvbic.dcoe.mil/dod-worldwide-numbers-tbi
2. L. K. Lindquist, H. C. Love, and E. B. Elbogen, "Traumatic brain injury in Iraq and Afghanistan veterans: New results from a national random sample study," *The J. Neuropsychiatry Clin. Neurosci.*, vol. 29, no. 3, pp. 254–259, 2017.
3. A. Gregg, "Pentagon wants to spot illnesses by monitoring soldiers' smartphones," [Online]. Available: https://www.washingtonpost.com/business/capitalbusiness/the-pentagon-wants-to-spot-illnesses-by-monitoring- soldiers-smartphones/2018/04/ 13/5238a646-3f55-11e8-a7d1-e4efec6389f0_story.html
4. P. R. Center, "Mobile fact sheet, " [Online]. Available: https://www.pewresearch.org/internet/fact-sheet/mobile/
5. W. Gerych, E. Agu, and E. Rundensteiner, "Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach," in *Proc. IEEE 13th Int. Conf. Semantic Comput.*, 2019, pp. 124–127.
6. A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland, "Sensing the" health state" of a community," *IEEE Pervasive Comput.*, vol. 11, no. 4, pp. 36–45, Oct.-Dec. 2012.
7. M. M. Baig, H. GholamHosseini, and M. J. Connolly, "Mobile healthcare applications: System design review, critical issues and challenges," *Australas. Phys. Eng. Sci. Med.*, vol. 38, no. 1, pp. 23–38, 2015.

8. M. S. H. Aung *et al.*, "Leveraging multi-modal sensing for mobile health: A case review in chronic pain," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 5, pp. 962–974, Aug. 2016.
9. Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Comput.*, vol. 16, no. 4, pp. 62–74, Oct.-Dec. 2017.
10. Q. He and E. O. Agu, "Smartphone usage contexts and sensable patterns as predictors of future sedentary behaviors," in *Proc. IEEE Healthcare Innov. Point-Of-Care Technol. Conf.*, 2016, pp. 54–57.
11. D. Roggen *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sensing Syst.*, 2010, pp. 233–240.
12. H. Mansoor, W. Gerych, L. Buquicchio, K. Chandrasekaran, E. Agu, and E. Rundensteiner, "Delfi: Mislabelled human context detection using multi-feature similarity linking," in *Proc. IEEE Visualization Data Sci.*, 2019, pp. 11–19.
13. B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
14. M. Zeni, W. Zhang, E. Bignotti, A. Passerini, and F. Giunchiglia, "Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge," *Proc. Assoc. Comput. Mach. Interact. Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019, Art. no. 32. [Online]. Available: https://doi.org/10.1145/3314419
15. M. Dehghani, A. Severyn, S. Rothe, and J. Kamps, "Learning to learn from weak supervision by full supervision," in *Meta-Learn. NIPS Workshop*, 2017.
16. N. I. of Neurological Disorders and Stroke, "Tremor fact sheet," 2019. Accessed: Oct. 3, 2019. [Online]. Available: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Tremor-Fact-Sheet
17. DARPA. "Darpa wash baa". [Online]. Available: https://beta.sam.gov/opp/cfb9742c60d055931003e6386d98c044/view
18. A. Alajaji, W. Gerych, K. Chandrasekaran, L. Buquicchio, E. Agu, and E. Rundensteiner, "Deepcontext: Parameterized compatibility-based attention CNN for human context recognition," in *Proc. IEEE 14th Int. Conf. Semantic Comput.*, 2020, pp. 53–60.
19. Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

RELATED WORK REFERENCES

R1. J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3995–4001.
R2. V. Radu *et al.*, "Multimodal deep learning for activity and context recognition," *Assoc. Comput. Machinery J. Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–27, Jan. 2018.
R3. N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," Apr. 2016, *arXiv:1604.08880*.
R4. S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. F. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. 26th Int. Conf. World Wide Web*, 2016, pp. 351–360.
R5. F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115–250, Jan. 2016.
R6. A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.
R7. W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman, "Classification accuracies of physical activities using smartphone motion sensors," *J. Med. Internet Res.*, vol. 14, no. 5, 2012, Art. no. e130.
R8. X. Su, H. Tong, and P. Ji, "Activity recognition with smartphone sensors," *Tsinghua Sci. Technol.*, vol. 19, no. 3, pp. 235–249, 2014.
R9. J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.
R10. Y. Vaizman, N. Weibel, and G. Lanckriet, "Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification," *Proc. Assoc. Comput. Machinery Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–22, 2018.

R11. P. Casale, O. Pujol, and P. Radeva, "Personalization and user verification in wearable systems using biometric walking patterns," *Pers. Ubiquitous Comput.*, vol. 16, no. 5, pp. 563–580, 2012.
R12. D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Public domain dataset for human activity recognition using smartphones," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2013.
R13. D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, 2017, Art. no. 1101.
R14. Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Comput.*, vol. 16, no. 4, pp. 62–74, Oct.-Dec. 2017.
R15. H. Blunck *et al.*, "Activity recognition on smart devices: Dealing with diversity in the wild," *GetMobile: Mobile Comput. Commun.*, vol. 20, no. 1, p. 34–38, Jul. 2016.
R16. F. Mordelet and J.-P. Vert, "A bagging SVM to learn from positive and unlabeled examples," *Pattern Recognit. Lett.*, vol. 37, pp. 201–209, 2014.
R17. M. Zeng, T. Yu, X. Wang, L. T. Nguyen, O. J. Mengshoel, and I. Lane, "Semi-supervised convolutional neural networks for human activity recognition," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 522–529.
R18. K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, May 2020.
R19. A. Akbari and R. Jafari, "Transferring activity recognition models for new wearable sensors with deep generative domain adaptation," in *Proc. 18th Int. Conf. Inf. Process. Sensor Netw.*, 2019, pp. 85–96.
R20. V. Radu and M. Henne, "Vision2sensor: Knowledge transfer across sensing modalities for human activity recognition," in *Proc. Assoc. Comput. Machinery Int., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–21, 2019.

ABDULAZIZ ALAJAJI is currently working toward the Ph.D. degree in data science with the Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include developing machine learning algorithms for mobile health. Contact him at asalajaji@wpi.edu.

WALTER GERYCH is currently working toward the Ph.D. degree in data science with the Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include developing machine learning algorithms to learn from incompletely labeled data. Contact him at wgerych@wpi.edu.

LUKE BUQUICCHIO is currently working toward the Ph.D. degree in data science with the Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include developing machine learning algorithms to understand the emergence of new, unknown classes. Contact him at ljbuquicchio@wpi.edu.

KAVIN CHANDRASEKARAN is currently working toward the Ph.D. degree in data science with the Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include developing machine learning algorithms for recognizing complex human activities. Contact him at kchandrasekaran@wpi.edu.

HAMID MANSOOR is currently working toward the Ph.D. degree in computer science with the Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include building interactive data visualizations to analyze multivariate data. Contact him at hmansoor@wpi.edu.

EMMANUEL AGU is currently a professor with the Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA. His research interests include computer graphics, mobile computing, image analysis, and machine learning, especially applications in healthcare. Contact him at emmanuel@wpi.edu.

ELKE A. RUNDENSTEINER is currently a professor of computer science and is the founding Director of the interdisciplinary Data Science program with the Worcester Polytechnic Institute, Worcester, MA, USA. As an internationally recognized expert in big data analytics, her research spans data science, data stream analytics, machine learning, visual and computational big data infrastructures, and digital health. Contact her at rundenst@wpi.edu.