Bennett CC, et al. (2023) Enabling robotic pets to autonomously adapt their own behaviors to enhance therapeutic effects: A data-driven approach. *IEEE International Conference on Robot and Human Interactive Communication (ROMAN).* In Press.

Enabling robotic pets to autonomously adapt their own behaviors to enhance therapeutic effects: A data-driven approach*

Casey C. Bennett, *Member IEEE*, Selma Sabanovic, *Member IEEE*, Cedomir Stanojevic, Zachary Henkel, Seongcheol Kim, Jinjae Lee, Kenna Baugus, Jennifer A. Piatt, Janghoon Yu, Jiyeong Oh, Sawyer Collins, Cindy L. Bethel, *Member IEEE*

***Abstract*— Socially-assistive robots (SARs) hold significant potential to transform the management of chronic healthcare conditions (e.g. diabetes, Alzheimer's, dementia) outside the clinic walls. However doing so entails embedding such autonomous robots into people's daily lives and home living environments, which are deeply shaped by the cultural and geographic locations within which they are situated. That begs the question whether we can design autonomous interactive behaviors between SARs and humans based on universal machine learning (ML) and deep learning (DL) models of robotic sensor data that would work across such diverse environments? To investigate this, we conducted a long-term user study with 26 participants across two diverse locations (United States and South Korea) with SARs deployed in each user's home for several weeks. We collected robotic sensor data every second of every day, combined with sophisticated ecological momentary assessment (EMA) sampling techniques, to generate a large-scale dataset of over 270 million data points representing 173 hours of randomly-sampled naturalistic interaction data between the human and SAR. Models built on that data were capable of achieving nearly 84% accuracy for detecting specific interaction modalities (AUC 0.885) when trained/tested on the same location, though suffered significant performance drops when applied to a different location. Further analysis and participant interviews showed that was likely due to differences in home living environments in the US and Korea. The results suggest that our ability to create adaptable behaviors for robotic pets may be dependent on the human-robot interaction (HRI) data available for modeling.**

Keywords: Human Robot Interaction, Socially Assistive Robots, Machine Learning, Ecological Momentary Assessment, Cross-Cultural Robotics

*This research was supported by the research fund of Hanyang University (HY-2020) in Korea, as well as from the National Science Foundation grant (IIS-1900683) in the US.

Casey C. Bennett is with the Department of Intelligence Computing, Hanyang University, Seoul Korea 04763, and the School of Computing at DePaul University, Chicago USA (e-mail: cabennet@hanyang.ac.kr).

Selma Sabanovic and Sawyer Collings is with the School of Informatics, Computing, and Engineering at Indiana University (email: {selmas, sercolli}@indiana.edu

Cheda Stanojevic is with the Department of Recreation and Tourism Management at Clemson University, USA (email: cstanoj@clemson.edu)

Zachary Henkel, Kenna Baugus, and Cindy Bethel are with the Department of Mechanical Engineering at Mississippi State University, USA (email: zmh68, kbb269. CBethel)@ msstate.edu

Seongcheol Kim, Jinjae Lee, Janghoon Yu and Jiyeong Oh are also with the Department of Intelligence Computing, Hanyang University, Seoul Korea (email: {sckim219, jjlee93, jqdjhy, ojyhi010309}@hanyang.ac.kr)

Jennifer A. Piatt are with the School of Public Health, Indiana University, Bloomington USA 47408 (e-mail: jenpiatt@indiana.edu) .

I. Introduction

A. Background

A fundamental challenge in the widespread deployment of interactive robots into people's everyday living spaces (home, work, etc.) is designing those interactions in a way that is *meaningful* to the end user. Indeed, it is relatively easy to design interactions from the perspective of the designer, but understanding how users actually interact with our agents in the real-world when we are not there actively observing is more difficult [1,2]. This is of particular interest during human-robot interaction (HRI) with socially-assistive robots (SARs), where we have embodied robotic agents in users' homes equipped with an array of sensors that can collect data about the interaction and the environment within which it takes place. Such sensor data can be used to understand the interaction behaviors occurring in real-time by classifying multi-modal sensor patterns into discernible activities, with the aim of generating models for intelligent robot control [3]. Those autonomous robots could then serve various purposes, such as enhancing human health or assisting with long-term chronic conditions, e.g. dementia or Alzheimer's [4,5].

A common assumption in social robotics is that models of interactive behavior built in one geographic location will seamlessly transfer to another location, from one culture to another, with perhaps just a little adaptation [6,7]. However, it is still an open question whether that is the case, or whether behavioral models – e.g. machine learning (ML) or deep learning (DL) models – created in one location would in fact be sub-optimal in another. **Can we truly develop a single social robotic platform, running the same algorithms, and then use it in many different culturally-distinct geographic locations?** Would the interaction patterns even look the same at a second-by-second sensor data level? Would a robotic pet be truly adaptable using that approach?

Part of that assumption is out of necessity, as it is difficult to collect data in many different locations given the logistics and costs of replicating robots in multiple research labs and/or transporting them [8], let alone the challenges of conducting simultaneous identical human research trials across multiple countries [9]. Nevertheless, answering the above question requires that we run such HRI studies in real-world settings to generate naturalistic interaction data, followed by meticulous modeling using a variety of ML/DL techniques[1]. Yet doing so necessitates that we can sample such real-world HRI data in a rigorous yet replicable manner.

B. Modeling Real-World Human Robot Interactions

A number of techniques have previously been developed to understand real-world user interactions during HRI, in

healthcare and other settings. Those include participatory design and other recall-based methods to collect data post-hoc (after the interaction), e.g. diaries, phone calls, and other data-collection instruments at the end of each day/week [10-12]. However, such methods are limited by people's capacity to remember what occurred and their tendency to re-construct past events based on current perceptions (i.e. "recall bias") [13]. A different approach is to use ecological momentary assessment (EMA), which has been shown to be a powerful tool for monitoring everyday user behaviors by gathering real-time data via smartphones [14] as well as modeling health-related behaviors in particular [15,16]. EMA works by randomly sampling each user's behavior multiple times throughout the day over a period of time (days, weeks, months). More recent research has begun to combine interactive robots (e.g. SARs) with EMA techniques [17,18].

Prior work has also looked at using sensor data during HRI studies to detect certain features of the interaction, such as affect [19], pose estimation [20,21], and gesture recognition [3], among others. Many of these studies, however, were in lab settings, and a challenge remains for identifying free-form activity patterns in everyday life where the same "behavior modality" can manifest in slightly different ways at different times, and thus would appear differently in the sensor data. For example, talking can be performed quite differently depending on if one is shouting at someone across the room, versus if they are whispering to their new puppy on the couch. Or take eating – e.g. eating pasta versus eating popcorn are the same behavior modality, but quite different in practice in terms of how the hands and mouth are moved. This is a similar challenge seen with human activity detection based on mobile phones, where a particular behavior (e.g. walking) can look different in the sensor data depending on if the person is carrying the phone in their hand, on its side in their purse, upside down in their jacket pocket, etc. [22]. EMA (and related *ambulatory assessment* techniques) offers one potential strategy for addressing this, by capturing the real-world variation of various behavior modalities [23]. For example, many researchers are starting to investigate the use of such smartphone data for dementia, multiple sclerosis, and other neurological diseases in terms of how changes in mood and cognition (i.e. "brain health") impact keyboard typing dynamics [24,25]. However, thus far the existing research on combining EMA with social robots is limited to a handful of relevant papers [17,18,26,27].

C. Research Aims

The primary aim of this research is to address the above questions by conducting a long-term deployment of a SAR companion pet in user homes across diverse geographic locations (South Korea and the United States), while using sophisticated sampling techniques to produce a large-scale dataset of randomly-sampled naturalistic human interactions with the robot. The sampling combined real-time robotic sensor data collected every second of every day along with EMA interaction data collected via a smartphone app. The data was then modeled using multiple ML/DL techniques to compare differences across geographic locations.

More broadly, the long-term aim of this research is to explore integration of in-home robots into a larger healthcare-focused internet-of-things (IoT) ecosystems [28]. Indeed, the true potential of in-home robotic data may come via combination with data from other devices in user homes (smartphones, wearables, other smart home devices) [15,17]. In a healthcare sense, such an approach may enable us to have a more holistic view of a patient's health on an everyday basis outside the clinic walls. We return to this topic in the Discussion section.

Figure 1. Robot cat wearing sensor collar

II. METHODS

A. Overview

We conducted a 1-year-long in-home user study with a SAR companion pet between August 2021 and July 2022, recruiting 26 participants (13 Korean, 13 US). Each individual participated in the study for approximately 1 month, during which the robot was deployed in their homes for roughly 3 weeks, with follow-up interviews conducted afterwards. The goal was to study these users intensively over a longer period of time, rather than study many users briefly [29]. Due to technical hardware failures during deployment leading to partial data loss, we had to exclude 3 of those participants from analysis. This left us with a final sample of 23 participants (12 Korean, 11 US). During the deployment phase, the robotic companion pet was equipped with a custom-made sensor collar to detect interaction data in the vicinity of the robot, including light, sound, motion, and indoor environmental conditions (see Figure 1) [30]. Simultaneously, EMA data was collected about the types of interaction modalities occurring. In short, the collar data became our "features" while the EMA data became the "targets" for modeling. The feature list is shown in Table 1.

Sensor collar data was collected roughly 9 times per second, every minute of every day, across the three-week deployment period. That produced roughly 11.7 million data points per participant, for a total sensor dataset of over 270 million data points. Approximately 65% of the time participants indicated no interaction with the robot was occurring at the time the EMA prompt arrived. After integrating the EMA and sensor collar data, there were approximately 173 hours of randomly-sampled naturalistic interaction data with the SAR representing nearly 700 in-home interactions available for modeling.

B. Data Description

The study here was designed using a *convergent parallel mixed method* approach [31], which incorporated multiple types of data collection (both quantitative and qualitative)

TABLE I. FEATURE LIST

TABLE II. CATEGORY	Features	Description
Light & Sound Sensor	lightVal, audioVal	Raw values from light and sound sensors
Accelerometer	accX, accY, accZ	Motion amount from accelerometer in x, y (lateral) and z (up/down) directions
Rotation	arc	Amount of rotational motion during interaction
Orientation	orientation	The orientation of the robot at a given time, based in accelerometer readings
Orientation Category	Landscape Right, Landscape Left, Portrait Up, Portrait Down, Flat	Specific orientation categories detected, using accelerometer manufacturer-specified thresholds
Orientation Transitions	orient_shift	Frequency of detected transitions between orientation categories
Sound Category	Quiet, Moderate, Loud	Specific sound categories detected, using sound sensor manufacturer specified thresholds
Sound Transitions	Quiet-Moderate, Quiet-Loud, Moderate-Quiet, Moderate-Loud, Loud-Quiet, Loud-Moderate	Frequency of detected transitions between sound categories during interaction
Indoor Environmental Conditions	temp, humidity, pressure (air)	Raw values for indoor environmental conditions
Air Quality	IAQ, co2Equivalent, gasResistance, breathVocEquivalent	Raw values for indoor air quality

to both model the interactions and better understand the patterns the models detected. The study included 26 participants, 13 from South Korea and 13 from the United States. The participants were drawn from the general population aged 20-35 and living alone, approximately 70% of the sample was female. Though in previous SAR research, we did not detect any interaction gender differences [30].

For deployment, each participant was given a SAR, in this case the Hasbro Joy-For-All robotic therapy pet (https://joyforall.com) equipped with a robotic sensor collar (see Figure 1). Participants were able to choose either a dog or cat version, based on personal preference. The sensor collar was developed through a research collaboration between Mississippi State University, Indiana University, and Hanyang University, and includes sensors that can detect light, sound, movement, indoor air quality, and other environmental health data in the vicinity of the robot (see Table 1 above). This was an updated "V2" version of the collar, intended to enhance the capabilities over previous versions [17,30]. It is fabricated via custom 3D printed designs, then assembled by hand, see Figure S1 in the online Supplementary Material: https://tinyurl.com/yw546474

While sensor data was collected via the collars, self-reported interaction behavior modalities were collected simultaneously using an Expiwell EMA mobile app (https://www.expiwell.com/). The EMA app was setup to collect data about the interaction modality (the type of behavior) and proximity (whether the interaction occurred near/far to the robot), based on common interaction behaviors observed in prior research during SAR use in in-home settings [4,10,11,32]. The modalities included both active interactions (e.g. petting, talking, playing, moving location) and passive interactions (e.g. watching television/media, eating together with the robot). Using an EMA approach [17] participants were pinged via their smartphone roughly 5-7 times per day (randomized across waking hours) and asked to report all interactions with and around the robot during the previous 15 minutes. Those prompts consisted of a 7-question survey to assess their interactions with the SAR (SoREMA instrument), along with additional psychological assessment questions to gauge user perception and emotional response post-interaction (instrument is described in [17]). Approximately 2/3 of the time though, users reported no interaction behavior to be occurring, which is to be expected in real-world settings where users are not forced to interact with the robot. Participants also sometimes reported multiple modalities occurring during the same interaction period (on average roughly 2 modality types per interaction).

At the beginning of the deployment period, participants were given instructions on interacting with the robot, using the EMA app, and the different types of interaction modalities, as well as asked to provide baseline information about their typical daily waking/sleeping schedule for setting up the EMA pings. All interviews and forms were done in the participant's native language (English or Korean) and conducted by fluent speakers in the US or Korea. Though on the Korean side, all participants were required to have at least intermediate proficiency in English (equivalent to TOEIC level B1) or higher to be eligible to participate. The study was approved by the IRBs of Indiana University (US) and Hanyang University (South Korea).

C. Analysis Methods

The analysis in this paper is broken into 3 parts: 1) a general analysis of interaction patterns between the groups using descriptive statistics, 2) a machine learning & deep learning analysis, 3) a qualitative analysis of participant interview data.

In order to better understand the interaction patterns, we conducted an **interaction modality frequency analysis** as well as a feature selection analysis. For the former, we were particularly interested if there were differences in how the US and Korean participants interacted with the SAR companion pet in terms of different types of interaction modalities performed with different frequencies. For the latter, multiple types of feature selection were explored via the python

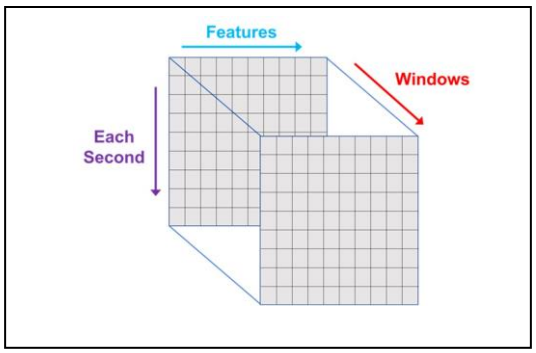
Scikit-Learn library, in order to attempt to identify which features were important for predicting *specific* modalities. That exploration included both wrapper-based and filter-based feature selection approaches [33].

Second, we analyzed the collected EMA and sensor collar data via **ML/DL modeling**. In short, the EMA data became the "targets" (i.e. interaction modalities) while the sensor data became the "features" for the ML and DL models. For simplicity, we collapsed the dataset into a series of binary classification predictions (e.g. petting vs. not petting) rather than attempt a complex multi-class classification problem. Due to target class imbalance, the data was re-balanced using SMOTE [34]. For predicting the EMA target, the feature data for that 15-minute time period was sliced into 15-second-long overlapping windows, with 50% overlap (similar to [35]). The way these features were handled depended on the type of modeling method. In general, the ML approaches calculated averages or percentages/frequencies for each feature across all the windows in the entire 15-minute interaction time period resulting a single row of data for each target, whereas the DL approaches utilized the smaller time windows directly so that each interaction was broken into many temporal slices. For DL, the data can be visualized as a multi-dimensional array, with a row representing approximately 100-150 milliseconds of data ("y" dimension), a column for each sensor data feature ("x" dimension), and each 15-second window being a third "z" dimension (see Figure 2 for a visual example).

ML approaches were performed using the python package Scikit-Learn (https://scikit-learn.org). Multiple modeling methods were attempted: Random Forest, Gradient Boosting, Neural Networks, and Support Vector Machines (SVM). Models were generally run using the default parameters in Scikit. Results were evaluated using 5-fold cross-validation based on accuracy and AUC (area under curve) metrics, following standard ML guidelines [36]. DL modeling was performed using the python package Keras (https://keras.io/), which is a deep learning library based on TensorFlow. To evaluate performance, 20% of the data was held out as a test set for each classification run. Multiple DL architectures were explored for comparison, which generally involved some combination of recurrent neural network (RNN) layers and convolutional neural network (CNN) layers. The idea was that the CNN could parse out "invariant representations" of pattern signatures occurring anywhere in the interaction, while the RNN could detect critical "sequences" of those patterns over time. We also compared two types of RNN layers: gated recurrent units (GRU) and Long-Short Term Memory (LSTM). After experimentation, the optimal unit size for those RNN layers was determined to be around 200, while the optimal CNN layers were found to have filter size of 26 with kernel size of 8. We also experimented with different numbers of layers, though we found that adding more complexity beyond just a few layers did not necessarily improve model performance in this case.

Finally, we performed a **qualitative analysis** of participant interviews. These were first coded by two independent coders using the Atlas TI software (https://atlasti.com/), using a coding scheme developed for the project that included a hierarchy of codes. The top level

Figure 2. Keras Data Input (described in-text)



of the hierarchy ("Code Group 1") distinguished comments related to the robot/collar, the EMA app, and the experiment itself. Below those top-level codes was a second level with codes for design, interactions, alerts, incentives, challenges, charging/battery issues, and desire for leaderboards or other types of gamification with the robot. For reference, the full code hierarchy is provided in the online Supplementary Material (Tables S6 and S7). The codes in the second level ("Code Group 2") were further broken into positive, negative, or suggestions for improvement. Interrater reliability between the two coders was calculated as 0.67, implying moderate agreement. After coding, the resulting data was analyzed in multiple ways. That included code occurrence frequencies, keyword analysis, and TF-IDF cosine similarities. We also conducted several t-test comparisons (two-way independent samples) between the Korean and US participant data for those analyses to detect any significant differences, which are described in the Results section.

III. Results

A. ML/DL Modeling Results

Our primary analysis was a comparison of ML/DL models for SAR interactions built based on either the Korean or US data separately (single-location) versus across locations or the entire dataset combined, in order to understand what might happen if a robotic pet built for users in one geographic locale was utilized in another locale. To do so, we evaluated five scenarios where we trained and then tested the ML/DL models on different datasets:

1) *Train on Korean data, test on Korean data (KOR Only)*
2) *Train on US data, test on US data (US only)*
3) *Train on Korean data, test on US data (Train KOR / Test US)*
4) *Train on US data, test on Korean data (Train US / Test KOR)*
5) *All data combined as if one dataset, then split for training/testing (US-KOR combined)*

Results can be seen in Table 2, with the DL models generally outperforming the ML models. For brevity here, we show only the best performing DL model (CNN+LSTM) and ML model (Random Forests), but more details can be found in the online Supplementary Material. The DL models worked the best on the Korean only data (scenario #1), achieving nearly 84% accuracy across all modalities (with

TABLE II. MODELING SCENARIO RESULTS (ACC=ACCURACY)

	ML		DL	
Dataset	**Acc**	**AUC**	**Acc**	**AUC**
KOR only	75.5	0.8327	83.6	0.8847
US only	73.8	0.8292	81.4	0.8501
Train KOR / Test US	65.0	0.5201	68.6	0.6969
Train US / Test KOR	66.2	0.4903	61.3	0.6212
US-KOR combined	67.5	0.7450	74.9	0.7919

AUC of 0.885). The US only data performed slightly less (scenario #2), but still achieved accuracy and AUC in the mid-80s. In contrast, models trained on one geographic locale then applied to another (scenario #3 and #4) did not work, exhibiting significant performance drops that would likely make them unusable for real-world applications.

When combining all the Korean and US data together as one dataset (scenario #5), we found that the DL models fell somewhere in between the single-location and cross-location scenarios. They had reduced performance compared to the former, though did perform better than the latter. The ML models in scenario #5 still performed similarly to the cross-location scenarios, however.

The primary takeaway from all this is that it appears that simply collecting HRI interaction data in one location to generate universal behavioral models for in-home robotic pets may not be a successful strategy. This is not entirely surprising, as the home living environments and lifestyles may be quite different in many cultures, e.g. Korea and the United States, which would in turn lead to differences in the sensor patterns of various interaction modalities with SARs. That suggests that we would need to sample data from multiple geographic locations in order to build a composite dataset that captures a variety of idiosyncratic patterns for modeling purposes. Or alternatively we would need to create models for each specific cultural environment. That lies in direct contrast to the notion of cultural homophily, i.e. attempting to simply adapt the same robot to different cultures, which some HRI researchers (including ourselves unfortunately) have argued for in the past [6,7].

Additionally, we evaluated the performance of those above-mentioned models on each individual interaction modality. Results can be seen in Table 3 for the KOR only dataset (scenario #1), with additional results for the US only and Combined datasets in online Supplementary tables S1 and S2, as well as different modeling methods in tables S3-S5. The DL models were obviously much more consistent in their performance across modalities in all scenarios, which likely indicates there is a significant temporal pattern to the interactions (which the recurrent layers of the DL can detect, but the ML cannot). However, one can also see that all the models struggled in particular with the Listening TV/Media modality, especially on the Korean side. We note that in post-deployment interviews, the Korean participants indicated that they were often watching YouTube or other media on their phones, and that due to the small living spaces in Korea were usually wearing headphones. That likely would cause problems for the SAR companion pet, as it can only hear ambient sounds.

TABLE III. PERFORMANCE BY MODALITY (KOR ONLY DATASET)

			ML		DL	
Modality	**Interaction Count**	**Total Interaction Time (min)**	**Acc**	**AUC**	**Acc**	**AUC**
Petting	116	1740	66.4	0.7901	76.5	0.8317
Talking	39	585	77.3	0.8725	84.9	0.8753
Playing	23	345	90.2	0.9878	96.1	0.9833
Listening TV/Media	118	1770	51.1	0.5555	72.4	0.7663
Eating/Cooking	33	495	81.3	0.8960	81.3	0.8997
Moving It	27	405	86.9	0.9540	90.1	0.9520
Average	59.3	890	75.5	0.8327	83.6	0.8847

TABLE IV. MODALITY INTERACTION FREQUENCY ANALYSIS. EACH ROW TOTALS 100%

Place	**Petting**	**Talking**	**Playing**	**Listening TV / Media**	**Eating / Cooking**	**Moved It**
Korea	31.9%	10.7%	5.8%	32.4%	9.1%	7.4%
US	34.7%	16.0%	4.0%	16.0%	7.1%	21.8%

B. Interaction Frequency and Feature Selection Results

To understand more about factors driving the patterns seen in the modeling results in the previous section, we undertook an interaction modality frequency analysis and a feature selection analysis, comparing the Korean and US samples. The results of the interaction frequency comparison can be seen in Table 4. Most of the interactions were performed with similar frequencies across the Korean and US samples with two exceptions. First, Korean participants reported Listening TV/Media more often with the robot. The interview data showed that a common activity for Koreans was lying on their bed or sitting at a desk watching YouTube (as well other media like Netflix) with the robot beside them. They also reported wearing headphones during that activity, due to the small living spaces and lack of sound-proofing in many Korean apartments. Conversely, US participants reported moving the robot around more frequently (moving it to a different spot in the room, or carrying it from room to room). Obviously, the current generation of SAR companion pets cannot walk, so carrying the robot from place to place is a common activity. Again, our interpretation here was that this was due to the differences in home living environments between the US and Korea, with US homes much larger in floor space and a greater number of rooms on average (roughly twice the size, according to OECD data [37]).

We also conducted a feature selection analysis to see if certain sensor features in our dataset we related to the patterns of specific modalities. The feature selection was conducted using multiple approaches (see Methods section). The full table of results can be found in supplementary material table S8, but to briefly summarize here we found many commonalities across modalities but also some notable distinctions. For instance, we found that Talking, Listening TV/Media, and Eating/Cooking were indicated by louder sounds as well as frequent sound shifts between sound levels (e.g. Loud-Quiet). Talking and Moving were related to VOC and CO2 levels near the robot, which we theorized was

TABLE V. CODE OCCURRENCE PER PARTICIPANT, BY CULTURAL LOCALE (QUOTE-LEVEL). FOR SIGNIFICANCE: *<0.05, **<0.01, ***<0.001

	Occurrence Count (per participant)					
Code Name	**US**	**KOR**	**T stat**	***p*-val**	**Sign.**	**Odds Ratio**
Alert	3.00	3.82	1.31	0.2053		0.63
Challenges	1.60	2.09	1.04	0.3125		0.64
Charging	1.10	1.64	1.43	0.1717		0.56
Design	0.64	0.70	0.17	0.8693		0.96
Feelings	7.10	3.45	3.19	0.0075	**	2.28
Gamifications	0.60	1.00	1.13	0.2737		0.51
Incentivizing	1.70	1.45	0.62	0.5497		1.03
Interactions	2.20	1.73	0.82	0.4214		1.13
Leaderboard	1.00	0.82	0.58	0.5726		1.07

possibly related to human contact (e.g. breathing). Petting was indicated by a particular orientation (landscape left back, derived from the accelerometer), which we believe was related to the robot cat behaviors (specifically rolling over when stroked on its back). Both Playing and Petting were indicated by several motion and orientation features, along with the ambient light levels.

C. Qualitative Analysis

In order to develop a deeper understanding of the interaction patterns, each participant took part in a 45-minute interview after the deployment ended. Those interviews were coded by two independent coders, then analyzed by various methods (code co-occurrences, keyword analysis, etc.) to compare between the US and Korean participants. A visualization as a Sankey diagram of the code frequency and co-occurrence associations was first created (see Figure S2). That revealed there were two main thematic clusters: 1) negative feelings associated with the technology (robot/collar charging, design, alerts, challenges), and 2) more positive feelings associated with the interactive experience (including potential incentives and gamification). That said, we suspected there may be significant differences between the US and Korean participants, which have been observed in previous cross-cultural SAR research with prior versions of the robot and collar [30]. A statistical comparison of quote-level code occurrences per participant (Korean vs US, two-way independent samples t-test) can be seen in Table 5.

The only significant difference appeared in the Feelings code category, which US participants mentioned during the interviews twice as often as the Korean participants. However, as mentioned above, we know from previous research that Korean participants tended to be more critical of this kind of SAR technology, while US participants focused more on the interactive experience [30]. To test this with the current study interview data, we re-coded the data so there were two feeling-related clusters: one with quotes about Alerts, Challenges, Charging, and Design and one with Interactions, Incentivization, Gamifications, and Leaderboards. A t-test comparison of those clusters (two-way independent samples t-test) between the US and Korean participants detected a significant difference (p-value 0.045). A closer look at a few of the participant quotes highlights this. For instance, amongst the Korean participants many comments expressed discomfort with the technology and the frequent sounds it made:

- "*Actually, I didn't feel uncomfortable, but I think it would be nice to reform it [Collar] so that the user can feel a little more friendly. Like a real cat's necklace. Since this is a pet robot, I think it would be better.*"
- *"There was a cat's gesture … but I don't know if it's a reaction to my action or just [sic] a random reaction. When I leave it on, it sometimes reacts alone, like crying."*
- *"I was really worried about what should I do with the robot, at first, as I've never raised a pet before."*

In contrast, the US participants seemed to have more experience with raising pets in general, and the larger living spaces in the US [37] seemed more suitable for the current generation of SAR designs:

- *"Honestly, the study just kind of brought back memories and like when I got my puppy, so it just it felt good to like reflect on that. Or just to like look forward to having something to come home to I guess."*
- *"I enjoyed it [sic] if I had come home from my class or something I [can] look forward to being able to interact with the dog. Just being able to play with it and talk to it."*
- *"He actually came in really handy because my friend, she came to visit me and she's like super scared of thunderstorms and it was thundering real bad and she saw the cat, and I was like do you want to hold him. So, we turned him on and it actually really helped her."*

Finally, we would note that while many users expressed a desire to be able to see a summary of their own human-robot interactions and sensor collar data at the end of each day, the majority of participants were negative rather than positive when it came to being able to compare their data to other people (e.g. leaderboards). There were also many conflicting views on whether providing incentives or rewards based on the data (or other types of "gamification") was a good idea. For reference, additional information about the code hierarchy and definitions of various codes is included in the Supplementary Material, tables S6 and S7.

IV. DISCUSSION

A. Summary of Results

We conducted a long-term user study combining SARs in people's homes over several weeks with EMA sampling techniques in two culturally-distinct geographic locations (United States and South Korea) in order to understand whether ML/DL models built for social robots in one location would still work when applied to another location. Data was collected every second of every day from 26 participants across the locations, generating a large-scale dataset of over 270 million data points. Combined with the EMA sampling, this produced over 173 hours of randomly-sampled naturalistic human-robot interaction data with SARs in real

world environments, which was then used for modeling to detect various interaction behavior modalities.

Results showed that creating universal ML/DL models for SARs based on data from only one geographic location may not be a successful strategy. While models created in one location were successful when used in the same location (~84% accuracy, 0.885 AUC), when applied to a different location they suffered significant performance drops (into the mid-60s). This was true whether we applied models from the United States to South Korea, or vice versa. Conversely, models built from a combined dataset of interaction data sampled from multiple geographic locations fell somewhere in between the single-location and cross-location models (~75% accuracy). In general, the DL models outperformed the ML models across all scenarios, indicating that there may be some differences in the temporal patterns in how various behavior modalities are performed across geographic locations that subsequently show up in the sensor data, which are subsequently detectable by the recurrent DL layers.

To further understand those geographic and cultural differences, we also undertook an analysis into the human-robot interaction frequency of different modalities across locations, as well as conducting in-depth qualitative interviews with participants. Those revealed that some of the modeling results above are likely due to differences in home living environments between the US and Korea, which affect the way that different behavioral modalities are performed and human perceptions of SAR technology. These results have potentially significant implications for autonomous SARs deployed in the real-world, e.g. into patients' homes for healthcare purposes.

B. SARs as Socially-Situated Healthcare Tools

As mentioned in the Introduction section, the true power of SARs for healthcare applications outside the clinic walls may lie in ecosystems of interconnected technology to create a single IoT type system, combining in-home robots, smartphones, wearables, and other devices. To put it a different way, this sort of connected systems approach is akin to the "systems biology" approach in healthcare settings that seeks to combine different sources of data (e.g. genetic, clinical, behavioral, social determinants) to more holistically understand an individual person's health status [38,39]. That approach has radically altered the field for both clinicians and patients, through the extension and integration of new forms of multi-modal data beyond "traditional" clinical data [40]. Indeed, we would be remiss not to mention the potential of sensor data from social robots integrated into home IoT systems to provide useful information about people's everyday social and cognitive functioning back to other in-home devices or even to in-clinic electronic health record (EHR) systems.

Furthermore, such an IoT ecosystem may provide support for a richer set of interactions with embodied agents like SARs in user's homes. Extending the types of data available for the robot to information not available in the robot's *immediate* vicinity can obviously open up a broader array of detectable human activity patterns that may only be a vague signal in the onboard robotic sensors alone, as well as possibly enable triadic (or higher degree) interaction behaviors between the user and multiple devices (with the SAR being one) [41].

Perhaps less obvious is the potential to model the effects of human-robot interactions that go beyond the immediate moment in time, to see how the initial impact might create long-term ripple effects downstream for the user's life and health status. In order to fully understand such long-term ripple effects of SARs in a more data-driven manner likely will necessitate an IoT approach [28]. Such data could then be used to link interaction modalities to their longer-term consequences. In other words, if an in-home robot behaves in a particular way today, how will that impact the user's health in a week or month later? Might it even influence their interactions with other technology? This may also enable a path towards better personalization of embodied agents [42]. For instance, outside the scope of research one might imagine a "continuous EMA data collection system" developed as a smartphone mobile app or wearable device that could accompany SARs or other robots when deployed in user's homes, which could then modulate agent behavior autonomously through machine learning models based on IoT ecosystem data.

All of the above suggests that if we want to design SARs in a manner that maximizes their utility as a healthcare technology, particularly in in-home settings outside the clinic walls, then we need methods to better understand the socially-situated environments they will inhabit [43,44]. Given the differences from one location to another, or one culture to another, means that doing so will likely be a significant engineering challenge for future research.

ACKNOWLEDGMENT

This research was supported by the research fund of Hanyang University (HY-2020) in Korea, as well as the National Science Foundation grant (IIS-1900683) in the US.

REFERENCES

[1] M. Jung, and P. Hinds, "Robots in the wild: A time for more robust theories of human-robot interaction." *ACM Transactions on Human-Robot Interaction (THRI)*, vol.4, no.1, pp.1–5, 2018.

[2] I. Brinck and C. Balkenius, "Mutual recognition in human-robot interaction: A deflationary account." *Philosophy & Technology*, vol.33, no.1, 53-70, 2020.

[3] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational human-robot interaction." *Foundations and Trends in Robotics*, vol.4, no.2-3, 105–223, 2016.

[4] S. Sabanovic, C.C. Bennett, W.L. Chang, and L. Huber, "PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia." *IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, pp.1–6, 2013.

[5] A. Liang, I. Piroth, H. Robinson, B. MacDonald, M. Fisher, U.M. Nater, et al.. "A pilot randomized trial of a companion robot for people with dementia living in the community." *Journal of the American Medical Directors Association*, vol.18, no.10, pp. 871-878, 2017.

[6] V. Lim, M. Rooksby. and E.S. Cross, "Social robots on a global stage: establishing a role for culture during human–robot interaction." *International Journal of Social Robotics*, vol.13, no.6, pp.1307-1333, 2021.

[7] S. Sabanovic, C.C. Bennett, and H.R. Lee, "Towards culturally robust robots: A critical social perspective on robotics and culture." *Proceedings of the HRI Workshop on Culture-Aware Robotics*, 2014.

[8] D. Ullman, S. Aladia, and B.F. Malle, "Challenges and opportunities for replication science in HRI: A case study in human-robot trust."

ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 110-118, 2021.

[9] H.G. Eichler and F. Sweeney, "The evolution of clinical trials: Can we address the challenges of the future?" *Clinical Trials*, vol.15, no.1_suppl, pp.27-32, 2018.

[10] C.C. Bennett, S. Sabanovic, J.A. Piatt, S. Nagata, L. Eldridge, and N. Randall, "A robot a day keeps the blues away." *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 536-540, 2017.

[11] N. Randall, C.C. Bennett, S. Sabanovic, S. Nagata, L. Eldridge, S. Collins and J.A. Piatt, "More than just friends: In-home use and design recommendations for sensing socially assistive robots (SARs) by older adults with depression." *Paladyn, Journal of Behavioral Robotics*, vol.10, no.1, pp.237-255, 2019.

[12] L. Pu, W. Moyle, C. Jones, and M. Todorovic, "The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies." *The Gerontologist*, vol.59, no.1, pp.e37–e51, 2019.

[13] S. Lindsay, D. Jackson, G. Schofield, and P. Olivier, "Engaging older people using participatory design." *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1199–1208, 2012.

[14] C. Vesel H. Rashidisabet J. Zulueta, J.P. Stange, J. Duffecy, F. Hussain, et al., "Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study." *Journal of the American Medical Informatics Association*, vol.27, no.7, pp.1007–1018, 2020.

[15] C.C. Bennett, M.K. Ross, E. Baek, D. Kim, and A.D. Leow. Smartphone accelerometer data as a proxy for clinical data in modeling of bipolar disorder symptom trajectory. *Nature NPJ Digital Medicine*, vol.5, no.1, pp.181, 2022.

[16] R.E. Mastoras, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Kassie, T. Alsaadi, et al., "Touchscreen typing pattern analysis for remote detection of the depressive tendency." *Scientific Reports*, vol.9, no.1, pp.1-12, 2019.

[17] C.C. Bennett, C. Stanojevic, S. Sabanovic, J.A. Piatt, and S. Kim, "When no one is watching: Ecological momentary assessment to understand situated social robot use in healthcare." *ACM International Conference on Human-Agent Interaction (HAI)*, pp. 245-251, 2021.

[18] E.A. Björling, E. Rose, A. Davidson, R. Ren, and D. Wong, "Can we keep him forever? Teens' engagement and desire for emotional connection with a social robot." *International Journal of Social Robotics*, vol.12, no.1, pp.65-77, 2020.

[19] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction." *Pattern Analysis and Applications*, vol.9, no.1, pp. 58-69, 2006.

[20] I. Kostavelis, M. Vasileiadis, E. Skartados, A. Kargakos, D. Giakoumis, C.S. Bouganis, et al., "Understanding of human behavior with a robotic agent through daily activity analysis." *International Journal of Social Robotics*, vol.11, no.3, pp.437–462, 2019.

[21] A. Chrungoo, S.S. Manimaran, and B. Ravindran, "Activity recognition for natural human robot interaction." *International Conference on Social Robotics (ICSR)*, pp. 84–94, 2014.

[22] M. Straczkiewicz, P. James, and J.P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research." *Nature NPJ Digital Medicine*, vol.4, no.1, pp.1-15, 2021.

[23] S. Shiffman A.A. Stone, and M.R. Hufford, "Ecological momentary assessment." *Annual Review of Clinical Psychology*, vol.4, no.1-32 2008.

[24] V.P. Cornet and R.J. Holden, "Systematic review of smartphone-based passive sensing for health and wellbeing." *Journal of Biomedical Informatics*, vol.77, pp.120-132, 2018.

[25] A.M. Pellegrini, E.J. Huang, P.C. Staples, K.L. Hart, J.M. Lorme, H.E. Brown, et al., "Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort." *Brain and Behavior*, vol.12, no.2, pp.e02077, 2022.

[26] E.J. Rose, E. Björling, and M. Cakmak, "Participatory design with teens: a social robot design challenge." *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pp. 604-609, 2019.

[27] M. Kim, T. Kwon, and K. Kim, "Can human–robot interaction promote the same depth of social information processing as human–human interaction?" *International Journal of Social Robotics*, vol.10, no.1, pp.33–42, 2018.

[28] M.A. Al-Taee, W. Al-Nuaimy, Z.J. Muhsin, and A. Al-Ataby, "Robot assistant in management of diabetes in children based on the Internet of things." *IEEE Internet of Things Journal*, vol.4, no.2, pp.437–445, 2016.

[29] C. Stanojevic, L.A. Eldridge, J.L. McIntire, A. Bowen, S. Dawson, B. McCormick, et al., "Employing an international research collaboration framework to pilot test an evidence-based recreational therapy program." *Therapeutic Recreation Journal*, vol.56, no.1, pp.1-16, 2022.

[30] C.C. Bennett, C. Stanojevic, S. Kim, S. Sabanovic, J. Lee, J.A. Piatt, J. Yu, J. Oh, "Comparison of in-home robotic companion pet use in South Korea and the United States: A case study." *9th IEEE International Conference on Biomedical Robotics & Biomechatronics (BIOROB)*, pp. 1-7, 2022.

[31] J.W. Creswell and J.D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed methods Approaches*. Thousand Oaks, CA, USA: Sage Publications, 2017.

[32] S. Collins, S. Sabanovic, M. Fraune, N. Randall, L. Eldridge, J.A. Piatt, et al. "Sensing companions: Potential clinical uses of robot sensor data for home care of older adults with depression." *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 89-90, 2018.

[33] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, "A review of microarray datasets and applied feature selection methods." *Information Sciences*, vol.282, pp.111-135, 2014.

[34] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, vol.16, pp.321–357, 2002.

[35] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition." *IEEE Access*, vol.5, pp.3095–3110, 2017.

[36] J. Siebert, L. Joeckel, J. Heidrich, K. Nakamichi, K. Ohashi, I. Namba, et al., "Towards guidelines for assessing qualities of machine learning systems." *International Conference on the Quality of Information and Communications Technology (QUATIC)*, pp. 17-31, 2020.

[37] Organization for Economic Cooperation and Development, *OECD Family Database, HC2.1. Living Space*. Paris, France: OECD, 2021. https://www.oecd.org/els/family/HC2-1-Living-space.pdf

[38] Y.S. Fraiman and M.H. Wojcik, "The influence of social determinants of health on the genetic diagnostic odyssey: who remains undiagnosed, why, and to what effect?" *Pediatric Research*, vol.89, no.2, pp.295-300, 2021.

[39] J.F. Figueroa, A.B. Frakt, and A.K. Jha, "Addressing social determinants of health: time for a polysocial risk score." *Journal of the American Medical Association*, vol.323, no.16, pp.1553-1554, 2020.

[40] A.L. Silva de Lima, T. Hahn, L.J. Evers, N.M. De Vries, E. Cohen, M. Afek, et al., "Feasibility of large-scale deployment of multiple wearable sensors in Parkinson's disease." *PLOS One*, vol.12, no.12, pp.e018916, 2017.

[41] A. Ciortea, O. Boissier, A. Zimmermann, A.M. Florea, "Give agents some REST: A resource-oriented abstraction layer for internet- scale agent environments." *International Conference on Autonomous Agents and Multiagent System (AAMAS)*, vol.17, pp.1502–1504, 2017.

[42] J. Fan, L.C. Mion, L. Beuscher, A. Ullal, P.A. Newhouse, and N. Sarkar. "SAR-connect: a socially assistive robotic system to support activity and social engagement of older adults." *IEEE Transactions on Robotics*, vol.38, no.2, pp.1250-1269, 2021.

[43] H.R. Lee, S. Sabanovic, W.L. Chang, S. Nagata, J. Piatt, C.C. Bennett, and D. Hakken, "Steps toward participatory design of social robots: Mutual learning with older adults with depression." *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp.244–253, 2017.

[44] Y.S. Sefidgar, T. Weng, H. Harvey, S. Elliot, and M. Cakmak, "RobotIST: Interactive situated tangible robot programming." *Proceedings of the Symposium on Spatial User Interaction*, pp.141–149, 2018.