Intelligent Service Robotics
https://doi.org/10.1007/s11370-023-00496-0

ORIGINAL RESEARCH PAPER

Generative replay for multi-class modeling of human activities via sensor data from in-home robotic companion pets

Seongcheol Kim[1] · Casey C. Bennett[1,2] · Zachary Henkel[3] · Jinjae Lee[1] · Cedomir Stanojevic[4] · Kenna Baugus[3] · Cindy L. Bethel[3] · Jennifer A. Piatt[5] · Selma Šabanović[6]

Received: 3 June 2023 / Accepted: 14 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Deploying socially assistive robots (SARs) at home, such as robotic companion pets, can be useful for tracking behavioral and health-related changes in humans during lifestyle fluctuations over time, like those experienced during CoVID-19. However, a fundamental problem required when deploying autonomous agents such as SARs in people's everyday living spaces is understanding how users interact with those robots when not observed by researchers. One way to address that is to utilize novel modeling methods based on the robot's sensor data, combined with newer types of interaction evaluation such as ecological momentary assessment (EMA), to recognize behavior modalities. This paper presents such a study of human-specific behavior classification based on data collected through EMA and sensors attached onboard a SAR, which was deployed in user homes. Classification was conducted using *generative replay* models, which attempt to use encoding/decoding methods to emulate how human dreaming is thought to create perturbations of the same experience in order to learn more efficiently from less data. Both multi-class and binary classification were explored for comparison, using several types of generative replay (variational autoencoders, generative adversarial networks, semi-supervised GANs). The highest-performing binary model showed approximately 79% accuracy (AUC 0.83), though multi-class classification across all modalities only attained 33% accuracy (AUC 0.62, F1 0.25), despite various attempts to improve it. The paper here highlights the strengths and weaknesses of using generative replay for modeling during human–robot interaction in the real world and also suggests a number of research paths for future improvement.

Keywords Human–robot interaction · Machine learning · Generative adversarial network · Generative replay · Ecological momentary assessment · Human activity recognition

1 Introduction

Socially assistive robots (SARs) are one technology that can be used to better understand the impact of mental health on physical health in everyday life, such as by placing robotic companion pets in user homes. For instance, by detecting how users interact with the robot via sensors onboard the robot, we can try to link those interactive behaviors to health-related changes through machine learning techniques. However, a problem that arises is how to best handle sensor data from multiple types of sensors so that it can be discerned into identifiable activities that relate to real-time human–robot interaction (HRI) of interest. Obviously, the same human behavior is not performed in the *exact* same way all the time, but rather there are groups of similar behaviors that cluster into common *behavior modalities* which we care about detecting [1, 2]. To address that, various recall-based data col-

✉ Seongcheol Kim
sckim219@hanyang.ac.kr

✉ Casey C. Bennett
cabennet@hanyang.ac.kr

[1] Department of Intelligence Computing, Hanyang University, Seoul, Korea

[2] College of Computing & Digital Media, DePaul University, Chicago, IL, USA

[3] Department of Computer Science and Engineering, Mississipi State University, Starkville, MS, USA

[4] Department of Parks, Recreation, and Tourism Management, Clemson University, Clemson, SC, USA

[5] School of Public Health, Indiana University, Bloomington, IN, USA

[6] School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

Published online: 19 December 2023

lection techniques (daily diaries, telephone conversation, and other data collection methods) have been previously developed, but those methods are limited in that they are not always entirely accurate due to the fact that they depend on the user's memory and thus suffer from "recall bias" [1, 3–6].

One possible solution to this is to combine the robot's sensor data with newer types of interaction evaluation methods, such as *ecological momentary assessment* (EMA) [7]. EMA is a technique of random sampling multiple times a day to capture real-time human behavior, even when researchers are not directly observing user behavior. Prior studies have demonstrated that EMA is a powerful tool for monitoring daily user behavior by collecting real-time data via smartphones [8, 9]. As such, more recent work has attempted to apply EMA during HRI studies, in order to understand the interactions that occur between humans and robots in the real world in real-time [2, 10]. As mentioned above, random sampling of users' activities over extended periods of time through EMA is important because the manner in which a particular activity is performed during use of a SAR may vary over time, both for a single user as well across different users. Moreover, previous studies have found that carefully linking the variability and novelty of a robot's behavior with identification of the exact moments when human behavioral changes occur can contribute to understanding how SARs could be used to support sustainable health-related behavior change [11]. Therefore, EMA offers the potential to address the above problems by combining it with interactive robots to better understand HRI in-the-wild.

For SARs to interact with humans more autonomously in the future, appropriate modeling techniques will be required to identify activities of interest within a particular domain (e.g., healthcare) based on the data [12, 13]. Such modeling will help researchers design more advanced real-time interactive agents in the future, and furthermore, allow for the extension of SAR technology into the realm of Internet of Healthcare Things (health IOT), which envisions an ecosystem of devices in users homes and workspaces intended to contribute to human activity recognition and the classification thereof with the aim of improving people's everyday lives [14]. Robots like SARs can provide particular value in such health IOT settings, as the SARs have a dual-purpose role in that scenario where long-term it can also provide companionship and serve as a digital therapeutic (DTx) intervention device, along with collecting sensor data [15]. Indeed, the broader goal of many health IOT systems is to provide such DTx interventions, beyond just collecting data.

To that end, the goal of this paper is to explore various generative replay deep learning modeling methods for human activity recognition based on robotic sensor data from in-home settings during HRI, in order to evaluate the potential that such models have when applied to interactive autonomous agents in the real-world. Generative replay attempts to use encoding/decoding methods to first memorize and then reinforce learned patterns in the data, similar to how human learning is thought to be reinforced during sleep time [16]. To understand such potential, we explore various types of generative replay models, for both binary and multi-class classification, and compare their performance to previously reported results using other types of machine learning and deep learning models.

2 Related works

2.1 EMA and sensor data for human activity recognition

Prior studies have shown that EMA is a powerful tool for monitoring routine patient behavior via various random sampling and self-assessment techniques [17–20]. Related to that, there are also numerous prior studies that attempt to directly recognize human activities through sensing technology such as smartphones, otherwise referred to as *ambulatory assessment* [20–22]. Understanding human activity using sensor data requires identification of detectable behaviors that exist within the patterns of collected sensor data. Not all behaviors meet that requirement, either due to limitations in the sensing technology or because the choice of sensor suite by the technology designers (e.g., roboticists) [23]. In some prior HRI studies, researchers have attempted to address that issue by having participants perform certain behaviors in a controlled environment monitored by researchers [24], as well as by recording videos in people's everyday living spaces over a long period of time [25]. However, such approaches can be costly and cause serious privacy issues. Previous studies have found human participants are generally uncomfortable with cameras placed in their homes, making cameras an infeasible solution for real-world deployed robots [26, 27].

A potential solution to this problem is to utilize real-time self-reported data by the user (via EMA) as "ground truth" labels of behavior with the sensor data (e.g., participant behavior) as suggested in previous HRI research [10]. Similar "ground truth" EMA approaches have been used in recent research attempting to detect human activity with other technology, such as smartphones and wearables [9, 28, 29]. The aim in all those research applications is to move away from the difficulties with recall-based methods of data collection (see Sect. 1).

2.2 Temporal detection of human activities

Another topic of major concern when attempting to model real-world sensor data for human activity recognition is how to deal with the issue of "time". Traditionally, data collected from sensors such as accelerometers use a fixed-size *sliding time window* for preprocessing, which extracts features for human activity recognition models. Researchers choose an appropriate time window size (e.g., some number of seconds), and then the windows are slid along the temporal sequence of sensor data, with some degree of overlap between one window to the next (e.g., 50%) The challenge is that activities don't happen instantaneously, nor do all activities occur for the same lengths of time. For instance, sitting down in a chair is an event that typically unfolds over several seconds [30]. However, what may be an appropriate time window for one behavior (e.g., 2 s) may be too short for another. Or vice versa, if the time window is too long (e.g., 2 min), then a brief 1-s behavior may be nothing more than a tiny blip in the sensor data, completely undetectable.

Some prior studies have indicated that a window size of 1–2 s intervals provides the best balance between recognition speed and accuracy [31], though other studies have indicated longer window sizes (e.g., 10 s) [30, 32]. Furthermore, there are also studies that suggested the use of an adaptive time window for human activity recognition may be better if the human behaviors are periodic or quasi-periodic [33]. Likely, the optimal window size will depend on the behaviors of interest, which makes this a challenging issue [34]. Regardless, this issue is an area of active research in the field.

3 Methods

3.1 Data description

The study here included 12 participants in their 20's from South Korea (8 females, 4 male), who participated over a 3-week-long period by interacting with a SAR robotic pet in their own homes. All participants were recruited from the general population, with residence types that varied among single-person households, living with family members, and living in dormitories. Each participant was given a Joy-For-All robot pet (Fig. 1) from Hasbro, equipped with a separately built sensor collar (Fig. 2). 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. The sensor data was collected roughly 9 times per second, 24 h a day, which resulted in a total sensor dataset of nearly 150 million data points (roughly ∼12 million per participant). Since we built the sensor devices, that sampling rate was programmed into them, by design.

While sensor data were collected via the collars, self-reported interaction behavior modalities were collected simultaneously using an EMA mobile app (Expiwell, https://www.expiwell.com/). The experiment used a sampling method

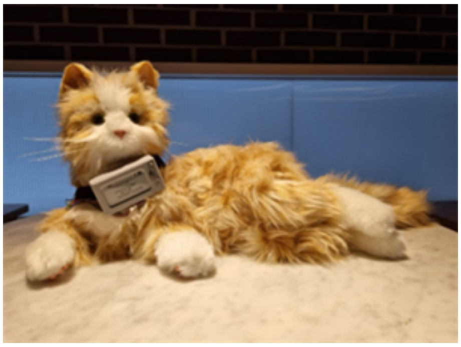

Fig. 1 SAR with attached sensor collar

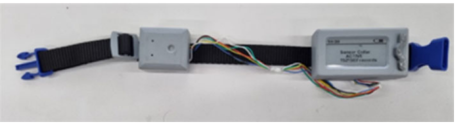

Fig. 2 Sensor collar

known as EMA to collect real-time data on interactions occurring between the robot pet and the participants [7], based on an approach previously developed specifically for use in HRI studies [10]. The EMA app was set to send notification prompts to users through their smartphones each day roughly 5–7 times, dividing the day (9 a.m. to 9 p.m.) into two hour segments, with the notification arriving randomly within a given time segment (or not at all). The aim was to capture realistic human behavior data with robots. Data were collected over a 16-day period, with pre- and post-questionnaires taking place in the days preceding the experiment and after. The full questionnaires, EMA prompts, and protocol can be found in [10].

The EMA prompts collected data about interaction behaviors (activity type) and proximity (whether the interaction occurred close to or far from the robot) over a 15-min period. Behavior modalities queried included direct interactions with the robot (petting, talking, playing) and indirect interactions (moving the robot, watching/listening to TV/YouTube/Radio, eating/cooking). 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.

This resulted in 364 samples of interactions across all modalities: petting, playing, moving the robot (from one location to another), talking, watching TV/radio (or other media such as YouTube), and eating/cooking. As each interaction represents a 15-min time period, the dataset represented roughly 91 h of total interaction time. As can be in Table 1, the modality data were imbalanced, so the training data were re-balanced using SMOTE [35]. These modalities represent the target class for modeling. A cleaned-up version of the dataset has been made publicly available in the Dryad

Table 1 Interaction modality counts and percentage

	Petting	Talking	TV/radio	Moved it	Playing	Eating/cooking
Count	116	39	118	27	15	16
Percentage	35.05%	11.78%	35.65%	8.16%	4.53%	4.83%

Table 2 Feature list

Category	Features	Description
Accelerometer	acc$X$, acc$Y$, acc$Z$	Motion amount from accelerometer in x, y (lateral) and z (up/down) directions
Rotation	Arc	Average amount of rotation motion during interaction
Light sensor	Full	Raw two-byte value reading from light sensor (visible + infrared)
Air quality sensor	Iaq, staticIaq, gasResistance, co2Equivalent, breathVocEquivalent	Raw average readings from air quality sensor
Environmental sensor	rawTemp	Raw average readings from air quality sensor (range: − 45 to 85 °C)
Environmental sensor	Pressure	Raw average readings from air quality sensor (range: 300 to 1100 hpa)
Environmental sensor	rawHumidity	Raw average readings from air quality sensor (range: 0 to 100%)
Indoor Air Quality Category	Good, Average, Little Bad, Bad, Worse, Very Bad	Percentage of time that specific Iaq categories were detected, using IAQ index manufacturer specified thresholds
Sound sensor	AudioLevel	Raw average readings from sound sensors
Sound category	Loud, Moderate, Quiet	Percentage of time that specific sound categories were detected, using sound sensor manufacturer specified thresholds
Orientation	Landscape Left Back, Landscape Left Front, Landscape Right Back, Landscape Right Front, Portrait Down Back, Portrait Down Front, Portrait Up Back, Portrait Up Front	Percentage of time that specific orientation categories were detected, using accelerometer manufacturer specified thresholds

Repository, both the Korean data used here and US dataset we collected using an identical procedure: https://doi.org/10.5061/dryad.tb2rbp078.

The features in our dataset were derived from the sensor collar, shown in Table 2. That included movement in the $x/y/z$ direction, rotational motion (arc), light/sound values, orientation of the robot, indoor air quality metrics, temperature, air pressure, humidity, CO_2 value, and volatile organic compound (VOC) rates related to human activities such as cooking.

The sensor data could not be used in its raw form because it was a continuous temporal sequence. To address this, a sliding time window was applied to divide the data into several short windows (see Sect. 2.2). The collar data for each interaction (a 15-min period) were split into 5-s windows with 50% overlapping (similar to [36]). This allowed each window to cover roughly 50 rows of sensor data, then move to the next, while still overlapping 50% with the prior window. That is to say, 25 rows of sensor data covered in the previous window are utilized again in the current window. Additionally, we applied some smoothing to the dataset via winsorization to eliminate any outliers or random deviations within individual features. As shown in Fig. 3, the final pre-processed input data took the form of a tensor (a multi-dimensional matrix) consisting of the set of features included in the collected sensor data (X-axis), with 50 data rows for each window (Y-axis), grouped into a series of 5-s overlapping time windows (Z-axis), similar to [12].

3.2 Generative replay models

In this study, both multi-class and binary classification were performed using generative replay models (aka "deep generative models"). Generative models make use of a strategy that attempts to learn resilient patterns of some phenomenon by approximating the distribution or characteristics of the

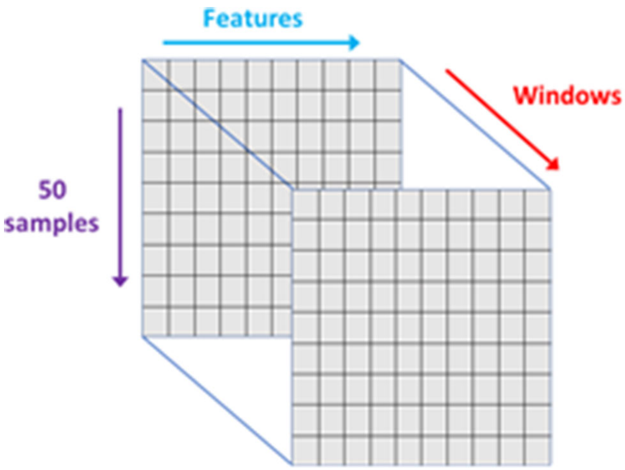

Fig. 3 Input data example

original dataset, then transferring those patterns to a small amount of data ("generator" or encoder) which can in turn be used to generate new data with small perturbations to prevent what is known as *catastrophic forgetting* during further training/testing as additional sensor input data arrives ("discriminator" or decoder) [37]. An example can be seen in Fig. 4, demonstrating how the process mirrors proposed hippocampus learning mechanisms in the human brain [16]. The aim is to increase the flexibility of learned patterns in data by creating variations of the same experience, which would allow us to learn more with less data. It is thought that this is similar to what the human brain is doing when dreaming (possibly explaining why dreams can seem so weird), and thus why the brain is so efficient at learning without needing millions of examples of something. Through that process, a trained generative model's generator and discriminator should hypothetically improve classification performance. Three different types of generative models are explored in this study: variable autoencoders (VAE) [38], generative adversarial networks (GAN) [39], and semi-supervised GANs (SGAN) [40].

In case of using VAE and GAN, they lack functionality for classification by themselves, so it is necessary to use a separate classification model structure for training the encoder and discriminator (i.e., transfer learning). Conversely, SGAN does not require transfer learning because classification training is possible within the SGAN itself. In this study, the training structure of the encoder and discriminator for VAE and GAN was approached into two ways. One was based on the architecture we have utilized in prior HRI studies [10, 12], which we henceforth refer to as the "CRNN-based" model. The other approach was based on a different architecture proposed by other researchers originally intended to classify human activities using data from wearable devices [13], which we will refer to as the "RCNN-based" model.

For comparison, we also used deep learning (DL) models in this study as a *baseline model*, which were drawn from our prior studies with a similar dataset [10, 12]. These were constructed using both convolutional neural networks (CNN) and recurrent neural networks (RNN) based on long-short term memory (LSTM) or gated recurrent units (GRU). The CNN and RNN layers were "stacked" to create DL architectures. The idea is that CNNs can characterize invariant representations of sensor data patterns that occur at any time during interactions, while RNNs can detect important sequences of those patterns over time. The CRNN-based and RCNN-based encoder/decoder training for the generative replay models mentioned in the previous paragraph is based on the same idea, except instead of training a model we are training the encoder/decoder. The primary difference between the CRNN-based and RCNN-based models above is whether the CNN or RNN layers occur first (i.e., the order of the layers). In other words, the primary difference relates to whether we should first try to detect invariant representations of sensor patterns that recur at different points in time, or whether we should first try to detect patterns across time.

Modeling in this study used Tensorflow via Keras (https://keras.io/), a Python-based deep learning library. In the case of multi-class classification, the six modalities described above were classified simultaneously, while for binary classification those six modalities were classified as a series of parallel binary predictions (e.g., petting or not petting). To evaluate the performance of the models, 20% of the data was held out from training as a test set.

4 Results

4.1 Baseline model

Table 3 shows the result of binary classification using deep learning models developed on a similar dataset (reported in our prior studies [10, 12]), applied to our new dataset here. Those models were based on combining RNN and CNN layers in a more traditional deep learning architecture, and obtained near 80% accuracy for predicting the same set of modalities (AUC 0.82). As mentioned in Sect. 3.2, we take this as our *baseline model* to compare with various generative replay models.

To establish a baseline for multi-class classification, we adapted the above model for multi-class prediction then re-ran it. The results are shown in Table 4 with an accuracy of only about 8 to 8.5%, which lower than the random chance probability of 16.7% (1/6) that could be obtained just by "guessing" the answer without any model. We also ran the baseline multi-classification using LSTM and GRU layers for the RNN portion as a check, but there was no significant difference in the baseline. A closer analysis revealed

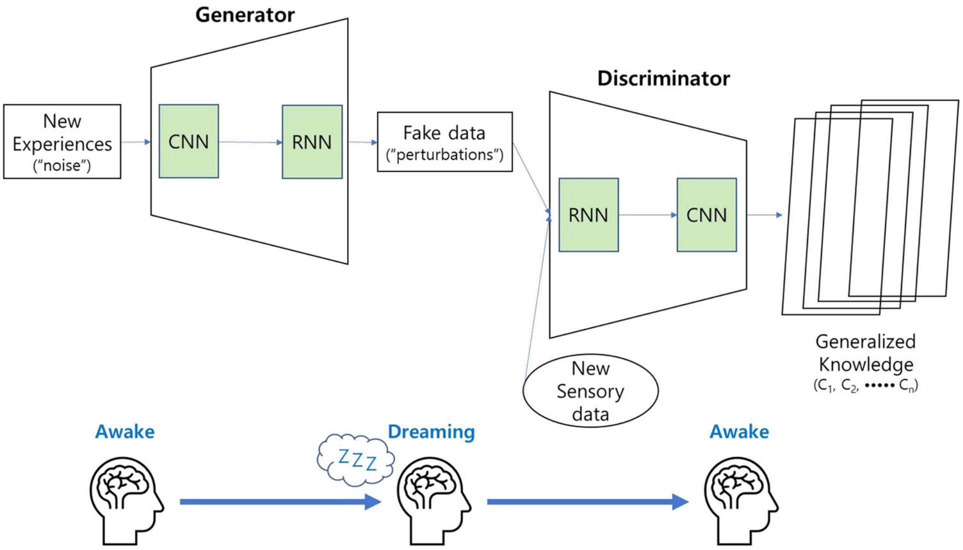

Fig. 4 Generate replay (GR) explained. Similar to the dreaming process during human sleep, GR generates perturbations of new recent experiences, which can be used to generalize knowledge when similar situations are encountered in the future

Table 3 Baseline binary classification results

Modality	Accuracy (%)	AUC
Petting	88.89	0.9406
Talking	79.57	0.8390
Playing	71.60	0.7827
TV/radio	63.79	0.6723
Eating/cooking	75.76	0.7782
Moving it	88.56	0.9184
Average	78.03	0.8219

Table 4 Baseline multi-class classification results

RNN model	Accuracy (%)	AUC	F1
GRU	8.12	0.1244	0.0429
LSTM	8.57	0.1472	0.0379

that the multi-class models struggled with the imbalanced distribution of the target class, which is a common problem in multi-class classification [41].

4.2 CRNN-based generative replay

4.2.1 Multi-class classification

We first evaluated generative replay using the CRNN-based model (see Sect. 3.2). Table 5 shows the results of multi-class classification. In general, all models showed a significant improvement over the baseline model in Table 4, with accuracies and other performance scores improved by a factor of about 2.5. However, the performance values are still suboptimal, and not really practical for real-world use with robotic pets. We do note that LSTM was slightly better than GRU on average in this case, which as we will see later was completely opposite with RCNN-based models.

Table 5 Multi-class classification results of CRNN-based models

Model	RNN type	Accuracy	AUC	F1
VAE	GRU	17.54	0.5941	0.1183
	LSTM	20.55	0.6264	0.1303
GAN	GRU	19.73	0.6113	0.1375
	LSTM	23.02	0.5389	0.1409
SGAN	GRU	18.91	0.4606	0.0993
	LSTM	21.65	0.5777	0.1462

4.2.2 Binary classification

Tables 6, 7, and 8 show the results of binary classification for the CRNN-based generative replay approach, using VAE, GAN, and SGAN models, respectively. Averse to multi-class classification above, it can be seen that the performance decreased in all cases compared to the baseline (Table 3). In particular, similar to baseline, there were certain modalities (Eating/Cooking, Listening to TV/Radio) which all models struggled with. More than likely, that is a "data issue" related to having the appropriate sensor suite onboard the robot to detect relevant modalities, rather than a modeling issue [10]. We return to this topic in the Discussion section.

4.3 RCNN-based generative replay

4.3.1 Multi-class classification

We next evaluated generative replay using the RCNN-based model (see Sect. 3.2). Results for multi-classification are

Table 6 Binary classification results of CRNN-based VAE

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	85.80	0.9351	82.61	0.9102
Talking	69.68	0.6958	82.26	0.8117
Playing	57.41	0.6256	59.26	0.6255
TV/radio	55.86	0.5790	51.72	0.5354
Eating/cooking	76.06	0.8151	87.88	0.9118
Moving it	73.73	0.7764	86.57	0.8657
Average	69.76	0.7378	75.05	0.7824

Table 7 Binary classification results of CRNN-based GAN

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	79.42	0.8716	80.87	0.8992
Talking	75.81	0.8121	70.97	0.7486
Playing	64.44	0.6563	72.22	0.7580
TV/radio	68.97	0.7186	58.97	0.6572
Eating/cooking	66.67	0.7306	63.94	0.6633
Moving it	71.04	0.7683	77.31	0.8220
Average	71.06	0.7596	70.71	0.7580

Table 8 Binary classification results of CRNN-based SGAN

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	81.81	0.8077	80.87	0.8872
Talking	77.58	0.8104	78.06	0.8359
Playing	64.07	0.6569	60.00	0.6382
TV/radio	62.76	0.6129	70.34	0.7245
Eating/cooking	79.24	0.8298	75.46	0.8085
Moving it	77.99	0.7981	81.79	0.8451
Average	73.91	0.7526	74.42	0.7899

shown in Table 9. Performance metrics were improved in all model types compared to baseline (Table 4). Averse to the CRNN multi-classification in Sect. 4.2.1, we note there was much greater variation in the results across models here, with accuracies ranging from 16 to 34%. Additionally, GRU seemed to perform better than LSTM with RCNN-based models, in contrast to the CRNN-based models where LSTM models performed better.

The SGAN using GRU was by far the best model of all the different multi-class classification approaches attempted in this study, achieving roughly 34% accuracy. That was an increase of a factor of 4× over the baseline model, and an increase of 2× over just random guessing. Regardless, our interpretation is that multi-class classification of interaction behaviors based on sensor data from in-home robotic pets is still sub-optimal for real world at the present time, though generative replay approaches do provide some significant advantages over more traditional deep learning approaches. The challenge is likely more of a sensor data issue, which will require experimentation with different sensor suites.

Table 9 Multi-class classification results of RCNN-based models

Model	RNN type	Accuracy (%)	AUC	F1
VAE	GRU	15.82	0.4740	0.1239
	LSTM	20.69	0.4755	0.1713
GAN	GRU	26.03	0.6169	0.1865
	LSTM	15.53	0.5015	0.0925
SGAN	GRU	33.79	0.6266	0.2588
	LSTM	27.53	0.5628	0.1849

4.3.2 Binary classification

Tables 10, 11, and 12 show the results of binary classification for the RCNN-based generative replay approach,

Table 10 Binary classification results of RCNN-based VAE

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	92.03	0.8733	91.39	0.9417
Talking	87.10	0.8565	80.00	0.8238
Playing	68.52	0.6772	63.89	0.7903
TV/radio	64.66	0.6791	57.76	0.5963
Eating/cooking	82.12	0.8626	75.00	0.7749
Moving it	74.10	0.7695	93.58	0.9321
Average	78.09	0.7863	76.94	0.8098

Table 11 Binary classification results of RCNN-based GAN

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	92.75	0.9430	92.75	0.9533
Talking	84.95	0.8923	77.96	0.8278
Playing	66.05	0.6953	65.43	0.6914
TV/radio	64.94	0.6969	64.37	0.6809
Eating/cooking	83.84	0.8845	72.22	0.7942
Moving it	75.12	0.7836	82.59	0.8870
Average	77.94	0.8159	75.89	0.8058

Table 12 Binary classification results of RCNN-based SGAN

Modality	GRU		LSTM	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Petting	87.51	0.8475	81.16	0.5572
Talking	77.42	0.8469	88.71	0.4882
Playing	67.59	0.7765	66.67	0.7877
TV/radio	62.07	0.7029	52.59	0.7347
Eating/cooking	88.64	0.9138	83.33	0.4054
Moving it	91.79	0.8965	73.13	0.5949
Average	79.17	0.8307	74.27	0.5947

using VAE, GAN, and SGAN models, respectively. We note that there was similar performance to the baseline model (Table 3) in this case, and much higher than the CRNN-based models. Similar to RCNN-based multi-class classification in Sect. 4.3.1, GRU was generally better than LSTM and the SGAN model type was the best performing of the generative replay models.

Similar to what was seen with the baseline model and CRNN-based models, there were particular modalities that caused the RCNN-based models difficulties. As such, these results concur with our previous point that that issue does not appear to be a modeling challenge, but rather a hardware issue related to the type of sensor data being collected onboard the robot. That was a consistent theme across all the modeling results.

5 Discussion

5.1 Summary of results

This study focused on evaluating deep generative replay models which could be applied to in-home robotic companion pets (e.g., SARs) so that sensor data onboard the robot could be used to recognize human activity in real time, which could then subsequently serve as a data-driven method for the robot to autonomously modulate its own interactive behaviors. For real-time interaction in users' homes and work environments, those are necessary capabilities for the creation of improved HRI systems [5, 10, 42]. Ideally, those capabilities would entail multi-class classification of many different types of interactions (e.g., playing, talking, eating), rather than only binary classification (e.g., talking vs. not talking, petting vs. not petting). However, multi-class classification is generally a harder challenge for most machine learning and deep learning modeling. In this study, we compared generative replay models with more traditional deep learning modeling approaches, using a dataset from 12 participants interacting with a robotic pet over several weeks in their home, comprising 91 h of total interaction time randomly sampled across the time period to obtain naturalistic interaction data.

The main takeaway from the results here was that using generative replay models for multi-class classification of robot sensor data provided a 4× improvement in performance over more traditional deep learning models. The RCNN-based SGAN models (using GRU) were the maximal-performing models. We can infer from that finding that there appears to be some value in using generative replay to emulate human dreaming via creating perturbations of the same data beyond the classic deep-learning approach. However, despite that, multi-class classification performance of interaction behaviors based on sensor data from in-home robotic pets was still sub-optimal for real-world use at the present time, regardless of modeling method. Beyond that, we also note that an RCNN-based architectures [13] outperformed the CRNN-based architectures. Initially, we thought the CRNN approach would perform better in a theoretical sense, i.e., first identifying invariant sensor patterns then looking for temporal sequences of those patterns. However, at least for our dataset here, it appears that first looking for distinctive temporal sequences then afterward identifying invariant patterns within those sequences (i.e., RCNN approach) is advantageous. That may be of use for future research on generative replay models in HRI.

Binary classification produced better results, though there were challenges with particular interaction modalities across all modeling types. More than likely, that is a "data issue" related to not having the appropriate sensor suite onboard the robot to detect relevant modalities [10, 12]. Without the correct data, no modeling method can detect patterns, so we think that it is more of a hardware issue at this point that will require the evaluation of different sensor suites to collect different sensor datasets. Those datasets could then be evaluated through a variety of feature selection and feature engineering methods to uncover which features are most useful for particular interaction modalities with robotic pets, as has been done with other types of interactive devices like smartphones [9, 21].

This study demonstrates the possibility of using deep generative replay models for recognizing specific human activities during HRI, and also points the direction for future research challenges.

5.2 Limitations and future work

There are number of limitations to the study here, mainly related to the dataset used and modeling choices made. We give some examples of those below.

In terms of the data, there were a number of issues with particular interaction modalities, which we noted in Sect. 4. For instance, prediction accuracy for TV/Radio behaviors (including YouTube viewing and other media) were notably lower than other modalities. A possible explanation for this is that, in the past, people generally did not use headphones when watching TV or other media. However, recently many people have a tendency to use headphones when watching media at home or while taking public transport (e.g., subway). Moreover, those effects may be more noticeable depending on location or population, e.g., urban vs. rural settings, older adults vs. young adults. Indeed, we realized this issue during a separate study comparing human participants interacting with robotic pets in the United States and Korea [12]. Those kinds of issues likely extend to other interaction modalities as well. As such, it remains for future research to evaluate other robotic sensor suites or types of data, or to perhaps even re-define what we think of as human activities given rapid technological changes in recent years [42, 43].

In terms of modeling choices, it is obviously necessary when undertaking this kind of study for researchers to decide on some set of parameters to explore, given that it is impossible to explore every possible option in a single study. For example, we note that the selected time window size (see Sect. 3.1) here may not necessarily have been optimal. Some prior studies have reported empirical results with a 1–2-s interval time window size as optimal [31], though other studies have reported larger window sizes of 10 s or more as optimal [30, 32]. In our case, it was judged that the data used in this study were complex and that the patterns underlying the interactive behaviors would sometimes last longer than just 1 s, so an intermediate 5-s interval was used. However, it is possible that window size was too large, or even too small. Alternatively, it may be necessary to use an adaptive time window rather than a fixed window size. Exploring the effects of these kinds of modeling choices also remains a challenge for future research.

In summary, real-time assessment of user activity using EMA and robotic sensor data holds the potential to serve as the basis for machine learning and deep learning models to classify human activities in in-home settings during HRI, as well as enable better interactive behaviors by social robots and other forms of AI. Understanding what users "actually do" with robots when researchers aren't there would allow the deployment of such models in order to allow robots like SARs to adjust their interactions in real time without the need to be "reprogrammed" by researchers or designers. Within in-home settings, the ultimate aim of such research is to extend the use of robotic devices for purposes like healthcare, e.g., as part of broader "Internet of Healthcare Things (IoHT)" ecosystems [44]. Work remains though to bring that concept to fruition.

Acknowledgements This work was supported by the research fund of Hanyang University (HY-2020) in Korea, as well as the National Science Foundation in the United States (Grant# IIS-1900683). We would also like to thank the numerous graduate and undergraduate research assistants who contributed to this work (Janghoon Yu, Jiyeong Oh, Zach Kauffman, Jonathan Gerth, Sawyer Collins).

Funding Funding was provided by Hanyang University (HY-2020) and National Science Foundation (IIS-1900683).

Data availability De-identified data may be made available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no conflicts or competing interests to declare related to this work.

References

1. Bennett CC, Sabanovic S, Piatt JA, Nagata S, Eldridge L, Randall N (2017) A robot a day keeps the blues away. In: 2017 IEEE international conference on healthcare informatics (ICHI). IEEE, pp 536–540
2. Bennett CC (2021) Evoking an intentional stance during human-agent social interaction: appearances can be deceiving. In: 2021 30th IEEE international conference on robot & human interactive communication (RO-MAN). IEEE, pp 362–368
3. Pu L, Moyle W, Jones C, Todorovic M (2019) The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. Gerontologist 59(1):37–51
4. Randall N, Bennett CC, Šabanović S, Nagata S, Eldridge L, Collins S, Piatt JA (2019) More than just friends: in-home use and design recommendations for sensing socially assistive robots (SARs) by older adults with depression. Paladyn J Behav Robot 10(1):237–255
5. Robinson H, MacDonald B, Broadbent E (2014) The role of healthcare robots for older people at home: a review. Int J Soc Robot 6(4):575–591
6. Lindsay S, Jackson D, Schofield G, Olivier P (2012) Engaging older people using participatory design. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1199–1208
7. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. Annu Rev Clin Psychol 4:1–32
8. Vesel C, Rashidisabet H, Zulueta J, Stange JP, Duffecy J, Hussain F, Piscitello A, Bark J, Langenecker SA, Young S et al (2020) Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A biaffect ios study. J Am Med Inform Assoc 27(7):1007–1018
9. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, McInnis M, Ajilore O, Nelson PC, Ryan K et al (2018) Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. J Med Internet Res 20(7):9775
10. Bennett CC, Stanojević, Č, Šabanović SA, Piatt J, Kim S (2021) When no one is watching: ecological momentary assessment to understand situated social robot use in healthcare. In: Proceedings of the 9th international conference on human–agent interaction, pp 245–251
11. Stanojević Č, Bennett CC, Šabanović S, Piatt JA, Kim S, Lee J (2022) Utilization of ema and the transtheoretical model of behavior change to prevent relapse during long-term socially assistive robot based interventions. In: Workshop on longitudinal social impacts of HRI over long-term deployments at the 2022 ACM/IEEE international conference on human robot interaction (LSI-HRI)
12. Bennett CC, Šabanović S, Kim SA, Piatt J, Lee J, Yu J, Oh J (in press) Comparison of in-home robotic companion pet use in South Korea and the united states: a case study. In: 9th IEEE international conference on biomedical robotics & biomechatronics (BIOROB)
13. Xia K, Huang J, Wang H (2020) LSTM-CNN architecture for human activity recognition. IEEE Access 8:56855–56866
14. Zhou X, Liang W, Kevin I, Wang K, Wang H, Yang LT, Jin Q (2020) Deep-learning-enhanced human activity recognition for internet of healthcare things. IEEE Internet Things J 7(7):6429–6438
15. Stanojevic C, Bennett CC, Sabanovic S, Collins S, Henkel KB, Henkel Z, Piatt JA (2023) Conceptualizing socially-assistive robots as a digital therapeutic tool in healthcare. Front Digit Health 5:15
16. van de Ven GM, Siegelmann HT, Tolias AS (2020) Brain-inspired replay for continual learning with artificial neural networks. Nat Commun 11(1):1–14
17. Eldridge L, Nagata S, Piatt J, Stanojevic C, Šabanović S, Bennett CC, Randall N et al (2020) Utilization of socially assistive robots in recreational therapy. Am J Recreat Ther 19(2):35–45
18. Vesel C, Rashidisabet H, Zulueta J, Stange JP, Duffecy J, Hussain F, Piscitello A, Bark J, Langenecker SA, Young S et al (2020) Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A biaffect ios study. J Am Med Inform Assoc 27(7):1007–1018
19. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker SA, McInnis M, Ajilore O, Nelson PC, Ryan K et al (2018) Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. J Med Internet Res 20(7):9775
20. Huckins JF, Wang W, Hedlund E, Rogers C, Nepal SK, Wu J, Obuchi M, Murphy EI, Meyer ML, Wagner DD et al (2020) Mental health and behavior of college students during the early phases of the covid-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 22(6):20185
21. Bennett CC, Ross MK, Baek E, Kim D, Leow AD (2022) Predicting clinically relevant changes in bipolar disorder outside the clinic walls based on pervasive technology interactions via smartphone typing dynamics. Pervasive Mob Comput 83:101598
22. Chen L, Hoey J, Nugent CD, Cook DJ, Yu Z (2012) Sensor-based activity recognition. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(6):790–808
23. Liu L, Dugas D, Cesari G, Siegwart R, Dubé R (2020) Robot navigation in crowded environments using deep reinforcement learning. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 5671–5677
24. Papagiannaki A, Zacharaki EI, Kalouris G, Kalogiannis S, Deltouzos K, Ellul J, Megalooikonomou V (2019) Recognizing physical activity of older people from wearable sensors and inconsistent data. Sensors 19(4):880
25. Jiang J, Pozza R, Gunnarsdóttir K, Gilbert N, Moessner K (2017) Using sensors to study home activities. J Sens Actuator Netw 6(4):32
26. Caine K, Sabanovic S, Carter M (2012) The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In: Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction, pp 343–350
27. Schulz TW, Herstad J (2018) Walking away from the robot: negotiating privacy with a robot. In: Electronic workshops in computing (eWiC), 2018. British Computer Society (BCS)

28. Bertz JW, Epstein DH, Preston KL (2018) Combining ecological momentary assessment with objective, ambulatory measures of behavior and physiology in substance-use research. Addict Behav 83:5–17
29. Huckins JF, Wang W, Hedlund E, Rogers C, Nepal SK, Wu J, Obuchi M, Murphy EI, Meyer ML, Wagner DD et al (2020) Mental health and behavior of college students during the early phases of the covid-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 22(6):20185
30. Lee K, Kwan M-P (2018) Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results. Comput Environ Urban Syst 67:124–131
31. Banos O, Galvez J-M, Damas M, Pomares H, Rojas I (2014) Window size impact in human activity recognition. Sensors 14(4):6474–6499
32. Niazi AH, Yazdansepas D, Gay JL, Maier FW, Ramaswamy L, Rasheed K, Buman MP (2017) Statistical analysis of window sizes and sampling rates in human activity recognition. In: HEALTHINF, pp 319–325
33. Sheng Z, Hailong C, Chuan J, Shaojun Z (2015) An adaptive time window method for human activity recognition. In: 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE). IEEE, pp 1188–1192
34. Ma C, Li W, Cao J, Du J, Li Q, Gravina R (2020) Adaptive sliding window based activity recognition for assisted livings. Inf. Fus. 53:55–65
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
36. Lee K, Kwan M-P (2018) Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results. Comput Environ Urban Syst 67:124–131
37. Shin H, Lee JK, Kim J, Kim J (2017) Continual learning with deep generative replay. Adv Neural Inf Process Syst 30
38. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114
39. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27
40. Odena A (2016) Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583
41. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M (2020) Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data 7(1):1–47
42. Jung M, Hinds P (2018) Robots in the wild: a time for more robust theories of human–robot interaction. ACM, New York
43. Brinck I, Balkenius C (2020) Mutual recognition in human–robot interaction: a deflationary account. Philos Technol 33(1):53–70
44. Zhou X, Liang W, Kevin I, Wang K, Wang H, Yang LT, Jin Q (2020) Deep-learning-enhanced human activity recognition for internet of healthcare things. IEEE Internet Things J 7(7):6429–6438

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.