IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 26, 2021, accepted January 4, 2022, date of publication January 18, 2022, date of current version January 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143990

Unsupervised Deep Learning to Detect Agitation From Videos in People With Dementia

SHEHROZ S. KHAN[1,2], PRATIK K. MISHRA[1,2], NIZWA JAVED[3], BING YE[2], KRISTINE NEWMAN[4], ALEX MIHAILIDIS[1,2], AND ANDREA IABONI[1,2]

[1]KITE—Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 2A2, Canada
[2]Institute of Biomedical Engineering, University of Toronto, Toronto, ON M5G 2A2, Canada
[3]The Centre for Vision Research, York University, Toronto, ON M5G 2A2, Canada
[4]Daphne Cockwell School of Nursing, Ryerson University, Toronto, ON M5G 2A2, Canada

Corresponding author: Shehroz S. Khan (shehroz.khan@uhn.ca)

This work was supported in part by SPARK Grant, Centre for Aging and Brain Health Innovation, Canada (2019–2020); in part by the Alzheimer's Society Research Program Catalyst Grant (ASRP 17-24); and in part by the Walter and Maria Schroeder Institute for Brain Innovation and Recovery. The work of Andrea Iaboni was supported by the Academic Scholars Award, Department of Psychiatry, University of Toronto.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by UHN REB under Approval No. 14-8483.

ABSTRACT Behavioural symptoms of dementia present a significant risk within Long Term Care (LTC) homes, which face difficulties supporting residents and monitoring their safety with limited staffing resources. Many LTC facilities have installed video surveillance systems in common areas that can help staff to observe residents; however, typically these video streams are not monitored. In this paper, we present the development of a computer vision algorithm to use these video streams to detect episodes of clinically important agitation in people with dementia. Given that episodes of agitation are rare in comparison to normal behaviours, we formulated this as an anomaly detection problem. This involves using the video camera to monitor the scene rather than tracking individuals. We developed a customized spatio-temporal convolution autoencoder that is trained on the normal behaviours and then identified agitation during testing as anomalous behaviour. We present a proof-of-concept using video data collected from a specialized dementia unit and annotated for agitation events. We trained the unsupervised neural network on approximately 24 hours of normal activities and tested on 11 hours of videos containing both normal activities and agitation events, and obtained an area under the curve of the receiver operating characteristic curve of 0.754. This research paves the way for leveraging existing surveillance infrastructure in LTC and other mental health settings to detect agitation or aggression, with the potential for improved health and safety.

INDEX TERMS Dementia, agitation, camera, long term care, autoencoder, deep learning, computer vision.

I. INTRODUCTION

Dementia is a disorder of progressive impairments of cognitive functions such as memory, language, and executive functioning, and can impact on insight, impulse control and judgement [1]. These cognitive changes contribute to changes in behaviour, which include behaviours that place the people with dementia (PwD) or those around them at risk, such as agitation [2]–[4]. In the later stages of the disease, PwD require supervision and support in their activities of daily living. Many PwD who need supervision and support live in long-term care (LTC) homes. In Canada, around one- third of PwD under 80 years live in a LTC home, this number increases to 42% for those who are above 80 years [5]. LTC home environments often suffer from a lack of staffing and financial resources that impacts on the quality of care of residents [6].

One application of technology in supporting the care of people with dementia is the use of sensors to detect episodes of clinically important behavioural symptoms [7], [8]. In our previous work, we have explored the use of wearable sensors to detect episodes of clinically important agitation [9], [10]. However, we discovered that it was difficult to use wearables in this population. One quarter of participants in this study

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

dissented to the use of the wearable and had to be withdrawn from the study. The wearable devices also need to be removed and re-applied to the participants on a daily basis to facilitate bathing, battery charging and data transfer [10], [11]. Thus, there are clear advantages to non- invasive sensors, such as video cameras, to be used for this purpose. Many LTC facilities have installed video surveillance to facilitate the digital monitoring of public spaces [12]. Video surveillance is zero-effort for staff and residents, and are designed to monitor the environment rather than the individual. However, most of the time, these video feeds are not monitored. Video data contains vital spatio-temporal information, which in conjunction with computer vision and artificial intelligence provides an opportunity to detect events or behaviours of risk from these video streams to support the care of PwD.

The other important challenge in developing algorithms for detecting agitation is the rarity of agitation episodes. The behavioural symptoms exhibited by PwD are episodic and they occur infrequently in comparison to normal activities [9]. A plausible approach to deal with this situation is anomaly detection [13] or one-class classification [14]. To detect anomalies, machine learning or deep learning models can be trained using only the data from normal observations (that are generally abundantly available) and these algorithms can then flag any significant deviations as anomalous behaviour [14]. Computer vision techniques have been successfully used in identifying anomalous behaviours in homes [15], crowded scenes [16] and public areas [17]. There has also been a lot of work in the general field of video based anomaly detection using deep learning methods [18], [19]. Nogas *et al.* [20] formulated the fall detection problem as an anomaly detection problem and used convolutional Long Short-Term Memory (LSTM) autoencoders to identify falls using videos collected from thermal cameras. It was observed that the convolutional LSTM autoencoders performed better than convolutional and deep autoencoders in detecting unseen falls. Further, the DeepFall framework [21] was proposed that used deep spatio-temporal convolutional autoencoders to learn spatial and temporal features from normal activities using video data collected by thermal and depth cameras. However, both the previous approaches were tested in semi-naturalistic conditions.

More naturalistic applications include the following recent works; CNNs and LSTM approach for classifying abnormal breathing events with 3D cameras [22], CNN and its variants and LSTM for fetal anomaly detection in ultrasound video scans [23], [24], CNN and LSTM method for aberrant epileptic seizures from videos [25], ResNet-based contrastive representation for abnormal otoscopy video sequences [26].

Two previous studies have used simulated data from videos to demonstrate possible approaches detecting behavioural symptoms in dementia. Fook et al [27] extracted hierarchical feature representation from the temporal segmentation maps of tracked patients using videos. They used probabilistic (Hidden Markov Model) and discriminative classifiers (Support Vector Machine) at different levels of hierarchies. However, the videos used in this analysis were simulated videos of a single person lying on a bed. The second study [28] used a dataset using Kinect camera from 10 participants who simulated various activities (hitting, pushing, throwing, tearing, kicking and wandering). Several joint-based features were extracted from this data, which were then combined using an ensemble learning method based on rotation forests.

In this paper, we make use of research video recordings from a Specialized Dementia Unit, using one camera view with approximately 35 hours of video data annotated with the behaviours of one research participant. We present this work as a proof-of-concept of an unsupervised deep learning approach, in which we train a spatio-temporal convolutional autoencoder only on the video recordings of normal behaviour of PwD and identify agitation during testing as anomalous behaviour.

II. METHODS

A. DESCRIPTION OF DATASET

Videos used in this study come from the Detecting Agitation study with 20 research participants on the Specialized Dementia Unit at TRI, located in Toronto, Canada and were collected between 2017 – 2019 [29]. Fifteen cameras were installed in public spaces (e.g., hallways, dining and recreation hall) of the unit. The Lorex model MCB7183 CCD bullet camera was used, having 700 TVL resolution with 960H optimized image sensor. Due to privacy concerns, the cameras were not installed in the bedrooms and washrooms of participating residents, and cameras only recorded between the hours of 0700 and 2300. The final dataset included annotated data for 17 participants (three participants were excluded due to lack of agitation events) with a mean age of 80.5 ± 9.1 years and 60% women. In total, 600 days' worth of video data were collected with 411 annotated agitation events. Thus, on an average, less than one agitation event per day was recorded. The duration of these agitation events varied significantly from 1 to 187 minutes. More than 230 hours of video data were manually reviewed to fine tune the agitation labels.

This study received research ethics approval (UHN REB #14-8483). Substitute decision makers provided written consent on behalf of the PwD for video recording. Written consent was also provided by the staff for video recording in the unit. Further written consent has been provided for publication of their images in video stills by all the staff and PwD appearing in the images shown in the paper. Any non-consenting individuals have been blackened out to hide their identities. For privacy and ethical reasons, the data is not publicly available.

B. DATA SELECTION AND ANNOTATION

For this proof-of-concept study, one participant was selected for analysis, based on the total number of hours of video

FIGURE 1. Normal events occurring in the dementia unit.

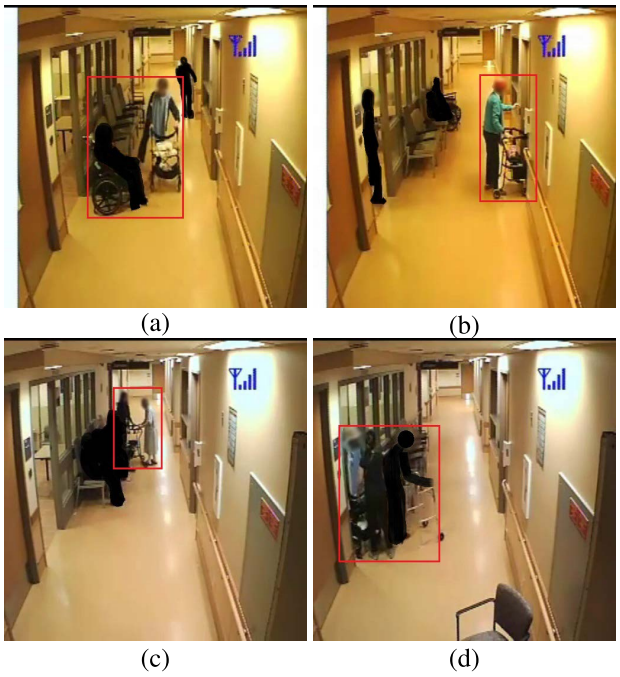


FIGURE 2. Agitation events annotated in the video data. The research participant is observed to (a) kick a co-patient. (b) bang on the nursing station door. (c) hit co-patient with walker. (d) push co-patient in the doorway. The bounding boxes are manually drawn to emphasize different agitation behaviours.

available (35 hours) and the number of agitation episodes captured within an area of the unit covered by a single camera. Using the clinical and research documentation for the days in question, a clinical research assistant reviewed the video 15 minutes before and after the time of documented episodes of agitation and annotated on a 30 seconds basis as to the presence or absence of agitation. The remainder of the video was considered to represent normal behaviour. For the purposes of this study, agitation was defined using the International Psychogeriatric Association definition [30], including excessive motor activity, verbal and physical aggression. In total, for this participant, 80 minutes out of the 35 hours were annotated as being part of an episode of agitation.

Figure 1 and 2 provide some example video frames depicting the normal and agitation events occurring in one of the hallway in the unit. It can be observed that for normal events, most of the time either the hallway is empty, or one more people walking around in the unit. However, in the case of anomalous (agitation) events, we can observe a patient kicking and pushing other patients and banging the door. The video data from this hallway is used to develop predictive models to detect agitation in one person with dementia.

C. DATA PREPROCESSING

A total of 35 hours of video data for the participant of interest was included and divided into training and test sets. The training set comprised of approximately 24 hours of video data, containing only normal events. The test set comprised of 11 hours of video data, which consisted of agitation events videos and 15 minutes of video data before and after them. The test set was divided into 30 second segments which were labelled as normal (0) or agitation (1) for the purpose of evaluating performance of the developed models.

The videos were sampled at 15 frames per second and the resultant frames were converted into grayscale, normalized and resized to 64×64 resolution. The frames were normalized by dividing the pixel values by 255 to keep them in the range [0, 1]. The gray scale conversion and resizing was done to reduce the computational cost in terms of trainable parameters. Finally, the frames were stacked to form non-overlapping 5 second windows, each window comprising of 75 frames. Therefore, each window has a dimension of $75 \times 64 \times 64$, where 75 denotes the temporal depth, and 64×64 denotes the spatial resolution of the frames. In the test video, if agitation behaviour was observed in a 30 second interval, then all the six 5-second windows were labeled as 1. The training videos were pre-processed to obtain a training set of 17355 normal activity windows, and was used to train the model. The test videos were pre-processed to obtain a test set of 7734 windows (6774 normal and 960 anomalous windows), and was used to test the model. Figure 3 presents the pipeline of this work.

III. SPATIO-TEMPORAL AUTOENCODER

To detect agitation as an anomaly, we use autoencoders to learn the underlying representation of normal activities from the video data of PwD and reconstruct the input video with minimum reconstruction error. After the training is accomplished, the autoencoder should be able to reconstruct an unseen normal video with low reconstruction error. However, in the case of an unseen anomalous event, a high reconstruction error is expected. Therefore, reconstruction

FIGURE 3. Pipeline for detecting agitation from videos.

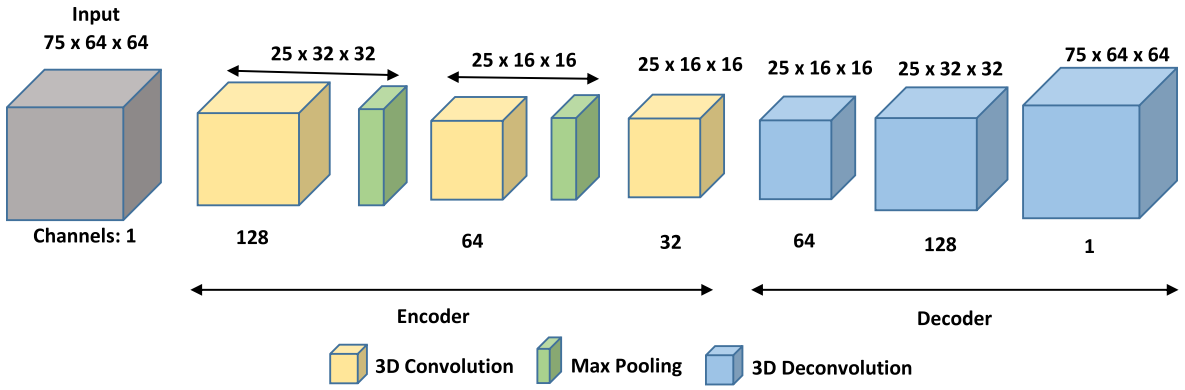


FIGURE 4. 3DCAE architecture to detect agitation in PwD as anomaly from the videos.

error can be used as a score to determine whether a test video sample is normal or anomalous (agitation in our case). Conventional autoencoders consisting of fully connected layers are not suitable for image/video data because of their inability to capture local spatial and temporal relationship in data [31]. 2D convolutional autoencoders can capture localized spatial features in images; however, they are unable to learn temporal features [32]. In this paper, we have used 3D convolutional autoencoder (3DCAE) [21] to learn spatio-temporal features from video data as an aide to learn normal scene and then use reconstruction error as a score to detect agitation as an anomaly.

The 3DCAE is composed of an encoder-decoder architecture (see Figure 4), which is adapted from the work of Nogas *et al.* [21] and customized for the problem of agitation detection. The input to the encoder consists of continuous frames stacked together forming a 3D window of dimension $75 \times 64 \times 64 \times 1$, where 75 denotes the temporal depth, 64×64 denotes the spatial resolution of the frames and 1 denotes the number of input channels, which is equivalent to gray scale image frames. The encoder further contains several convolution layers. The decoder operates in the reverse manner and reconstructs the input. The difference between the input and output frames is treated as a loss function / reconstruction error. The 3DCAE model learns the underlying representation of the data by minimizing the reconstruction error during training. At the test time, the model computes the reconstruction error for an unseen data, using the loss function and leverages the deviation in the error to identify instances of agitation as anomaly.

A. LOSS FUNCTION

We use mean squared error and gradient loss to calculate the reconstruction error and train the 3DCAE model. The mean squared error loss is used to minimize the difference in pixels between the input window frames I and the reconstructed window frames O as follows,

$$\mathcal{L}_{mse}(I, O) = \frac{1}{N_e} \sum_{l=1}^{W} \|I_l - O_l\|^2 \tag{1}$$

where, W represents the number of frames and is termed as the window size and N_e is the total number of pixels in a window. In the 3DCAE model, $W = 75$ and $N_e = 75 \times 64 \times 64 = 307200$.

The gradient loss [33] is used to sharpen the reconstructed images, and is defined as,

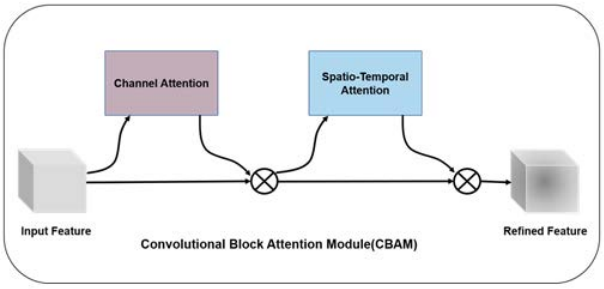
$$\begin{aligned}\mathcal{L}_{gd}(I, O) &= \sum_{l=1}^{W} \sum_{i=1}^{S} \sum_{j=1}^{S} |||O_{l,i,j} - O_{l,i-1,j}| - |I_{l,i,j} - I_{l,i-1,j}|||_1 \\ &\quad + \left\| |O_{l,i,j} - O_{l,i,j-1}| - |I_{l,i,j} - I_{l,i,j-1}| \right\|_1 \end{aligned} \tag{2}$$

where, S is the spatial size. In the 3DCAE model, $S = 64$.

Further, we combine both the losses to investigate their cumulative effect in training the 3DCAE model, and arrive at the following multi-objective combined loss,

$$\mathcal{L}_{msegd}(I, O) = \mathcal{L}_{mse} + \lambda \mathcal{L}_{gd} \tag{3}$$

where λ is a hyperparameter that has to be set empirically.



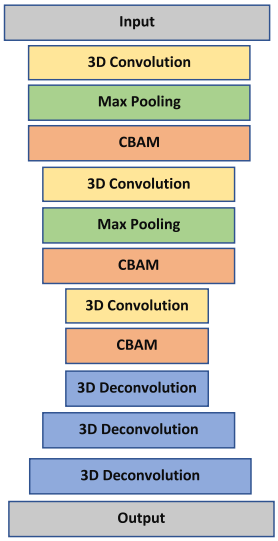


FIGURE 5. 3DCAE architecture with attention.

B. ATTENTION

In addition, we examined the efficacy of a Convolutional Block Attention Module (CBAM) [34] during the encoding phase to help the model learn better representation of video data. Originally CBAM was implemented on 2D convolutions for channel attention and spatial attention. Che and Peng [35], extended it to 3D convolutions for human action recognition tasks. 3D CBAM consists of two different modules – spatiotemporal and channel. The former focuses on spatial and temporal relation between features, whereas, the latter assigns weights to channels of a feature map based on their importance in reconstructing the input. In this paper, we explore the performance of 3D CBAM in unsupervised 3DCAE. These CBAM blocks are placed in between convolutional layers of the encoder as seen in Figure 5.

Given an input $I \in \mathbb{R}^{W\times H\times W_d\times C}$, where W is the window size, H is the height, W_d is the width and C is the channel size, the attention process in the CBAM module can be summarized as,

$$\begin{aligned} I' &= M_C(I) \otimes I \\ I'' &= M_{St}(I') \otimes I' \end{aligned} \tag{4}$$

where, $M_C \in \mathbb{R}^{1\times 1\times 1\times C}$ is the channel attention map and $M_{St} \in \mathbb{R}^{W\times H\times Wd\times 1}$ is the spatiotemporal attention map. The channel attention map and spatiotemporal attention map are calculated as,

$$\begin{aligned} M_C(I) &= \sigma(Conv3D(AvgPool(I)) \\ &\quad + Conv3D(MaxPool(I))) \\ M_{St}(I') &= \sigma(Conv3D(AvgPool(I'); MaxPool(I'))) \end{aligned} \tag{5}$$

IV. EXPERIMENTS AND RESULTS

The video data from one camera view was used, which included the participant of interest as well as many other patients, staff and visitors who also appear in the scene. We trained the 3DCAE model for 50 epochs. Adam optimizer was used with a learning rate of 0.001. The training batch size was fixed to 5, that is, each batch was made up of 5 windows. The reconstruction error was computed per window and used as an anomaly score. We use Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve as the evaluation metrics. The ROC curve helps to understand how well a classifier can generalize over different thresholds, while the PR curve highlights how relevant is a positive result from the classifier given the baseline probabilities of a problem. The anomaly score calculated on the test set was used to determine AUC, with agitation as the class of interest. The 3DCAE model was implemented in pytorch lightning [36].

A. DIFFERENT LOSS FUNCTIONS

To understand the effect of various loss functions on detecting agitation, we performed experiments on three different variants of 3DCAE model, namely, 3DCAE_mse, 3DCAE_gd, and 3DCAE_msegd. The text after the underscore represents that the 3DCAE model used the mean squared error loss (mse), gradient loss (gd) or their combined loss (msegd). To obtain the best value of hyperparameter λ in the combined mean squared and gradient loss, it is varied in the range $[0.01, 0.1, 1, 10, 100]$. Table 1 shows that $\lambda = 0.01$ gave the highest value of AUC of ROC and PR. Therefore, the value of λ for 3DCAE_msegd is kept as 0.01. We understand that reporting hyperparameters on testing set is not optimal; however, in the absence of a validation set it was considered plausible. The results of the experiments on all three variants is presented in Table 2. We observe that all the three models performed equivalently with 3DCAE_mse performing marginally better in terms of both AUC(ROC) and AUC(PR). With the addition of the attention model (CBAM) to the autoencoder, we did not observe any improvement to the baseline 3DCAE_mse model (Table 3, third row).

The baseline value for the PR curve is expressed as the ratio of the number of positive samples to the total number of samples. This value represents the behaviour of a random classifier. In the case of perfectly balanced classes, this value will be 0.5, whereas in the case of imbalanced classes, this values will be between 0 and 0.5. Therefore, the baseline

TABLE 1. Comparison of AUC scores for ROC and PR curves for different values of λ for 3DCAE_msegd.

λ	AUC (ROC)	AUC (PR)
0.01	0.749	0.254
0.1	0.741	0.248
1	0.747	0.253
10	0.748	0.254
100	0.747	0.252

TABLE 2. Comparison of AUC scores for ROC and PR curves for different variations of 3DCAE.

Model	AUC (ROC)	AUC (PR)
3DCAE_mse	0.754	0.259
3DCAE_gd	0.749	0.257
3DCAE_msegd	0.749	0.254

AUC value for our agitation dataset is $960/(6774 + 960) = 0.124$, assuming the events of agitation as the positive case. The low value of baseline is a result of the skewed data balance in case of agitation problem, as the episodes of agitation occur infrequently in comparison to normal activities. Among all the variants of 3DCAE, the minimum AUC(PR) score obtained is 0.254. Hence, all the 3DCAE models perform at least twice better than any random classifier in terms of AUC(PR) score for our agitation dataset

B. IQR ANALYSIS

The training video data is assumed to consist of normal events only. However, this data may as well contain few unreported agitation events or other behaviours of risks that could have been missed during the annotation process. These unreported events of risk may influence learning the 'normal' concept through 3DCAE. Therefore, some of the outliers in the normal events may be removed from the training data to facilitate learning of normal concept. Khan et al [37] proposed to remove such outliers in the training data using the inter-quartile range approach. Their general idea is to train a given model on the training data, perform inter-quartile analysis on the score obtained by running the trained model on the training samples and remove the samples with the highest/lowest scores. In our case, the model is 3DCAE_mse and the reconstruction error is used as the score.

Assuming, Q_1 as the lower quartile, Q_3 as the upper quartile, the inter-quartile range is $IQR = Q_3 - Q_1$. A sample P is considered as an outlier if,

$$P > (Q_3 + k \times IQR) \parallel P < (Q_1 - k \times IQR) \qquad (6)$$

where $\parallel$ denotes the logical OR operation and k is the rejection rate that represents the percentage of data points that are within the non-extreme limits. The extreme values of reconstruction error that represents the outliers in the training data can be removed and the model be trained on the remaining training samples.

In our experiments, we used $k = 1.5$ that accepts 99.3% of training data and remove the remaining samples from it. Figure 6 presents a box plot that shows the outliers based on the reconstruction errors of the training samples. A total of 940 samples were identified as outliers and removed from the training data.

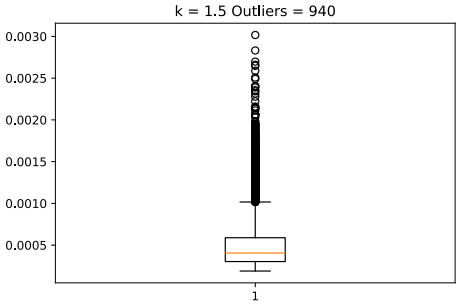


FIGURE 6. Box plot showing the outliers based on the reconstruction errors of the training samples of 3DCAE_mse.

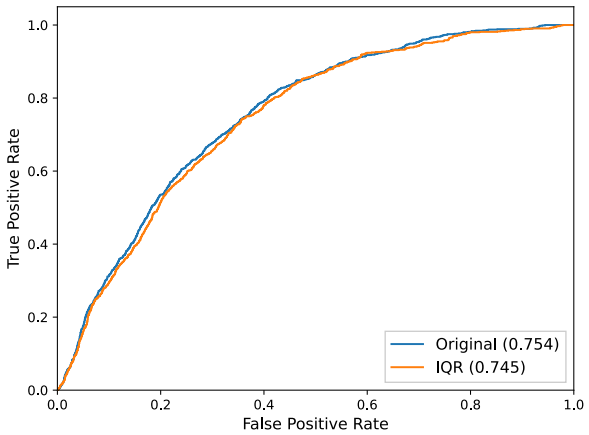
FIGURE 7. Some of the outliers found in training data as part of IQR analysis.

After the IQR analysis, the training set was reduced to 16415 windows. A few of the outlier frames are shown in Figure 7. Most of the outlier frames showed presence of large objects, such as trolleys or ladders and/or crowding of people in the scene. From a clinical perspective, these outliers in the training data can be useful information to identify potential triggers for agitation behaviours and preventing the occurrence of risky events happening in the unit.

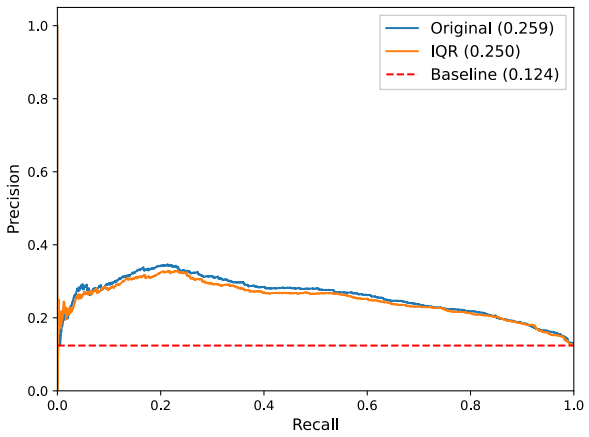
Table 3, second and fourth row provides the AUC scores and Figure 8 presents the ROC and PR plots for 3DCAE_mse before and after IQR analysis. In the figure, the values in parentheses in the labels refer to the AUC scores. As can be observed, there is no improvement in AUC scores after IQR analysis (Table 3), and the ROC and PR plots for the original training data and data obtained after IQR analysis are quite similar (Figure 8). This demonstrates that the 3DCAE model

TABLE 3. Comparison of AUC scores for different versions of 3DCAE_mse.

Model	AUC (ROC)	AUC (PR)
3DCAE_mse (derived from Table 2)	0.754	0.259
3DCAE_mse + CBAM	0.745	0.252
3DCAE_mse after IQR analysis	0.745	0.250
3DCAE_mse using 3 sec window	0.742	0.244
3DCAE_mse using 10 sec window	0.752	0.257



(a) ROC Plot.



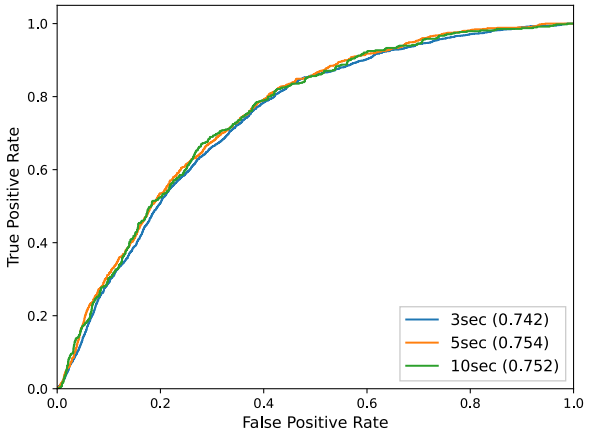
(b) PR Plot.

FIGURE 8. Comparison of AUC scores after IQR analysis for 3DCAE_mse.

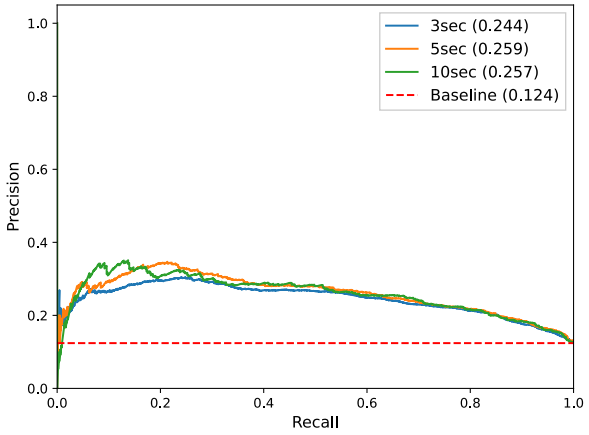
was robust to the presence of outliers in the training data and was able to learn normal behaviours occurring in the unit.

C. CHOICE OF TEMPORAL DEPTH

In the experiments shown in Section IV-A, the size of temporal window was chosen as 5 seconds. To investigate the effect of different sizes of windows on the performance of the 3DCAE for detecting agitation, we choose window sizes of 3 and 10 seconds. As discussed earlier, the test videos were labelled in intervals of 30 seconds. In order to



(a) ROC Plot.



(b) PR Plot.

FIGURE 9. Comparison of 3DCAE_mse model trained on varying input temporal depth.

obtain labels for 3 second windows, all the 3 second windows that fell under the same 30 seconds interval, were given the same label. Similar strategy was followed to label the 10 seconds windows. Table 3, fifth and sixth row, provides the AUC scores and Figure 9 presents the ROC and PR plots for 3DCAE_mse model for 3 and 10 seconds window sizes. In the figure, the values in parentheses in the labels refer to the AUC scores obtained for the corresponding window size. It can be observed that 5 and 10 seconds window size marginally performs better than the 3 second window size in terms of AUC(ROC) and AUC(PR) (Table 3 and Figure 9). However, larger window sizes require more computational resources.

V. DISCUSSION

In this paper, we demonstrate as a proof-of-concept that it is possible to detect agitation as a behavioural symptom of dementia in videos when formulated as an anomaly detection problem. We report the results from one camera view for one person with approximately 35 hours of annotated video data. We trained a spatio-temporal autoencoder on approximately

24 hours of normal video data and tested on 11 hours of data containing normal and agitation behaviours. Our approach resulted in an AUC for ROC of 0.754. To the best of our knowledge, this is one of the first research studies to present evidence on the use of video cameras in LTC for detecting agitation in real-life videos in a residential setting with PwD.

The IQR analysis demonstrated the large range of normal activities captured in the video in this setting, including some very busy and crowded scenes. Removing these scenes, however, did not improve upon the model performance, suggesting that the model is able to distinguish these scenes from anomalous events. The results suggested that the 3DCAE model with 5 second window size and mean squared error loss performed better than other parameter choices. The 3DCAE model is able to identify agitation with a 64 64 resolution, which may be unintelligible for naked human eyes. Therefore, this may provide a privacy protecting approach by partially/fully obfuscates the identity of the people in the scene. We also found lower values of AUC of PR curve indicating an increased amount of false positive rate during the testing. One reason could be the presence of other ''anomalous events'' in the testing test (besides agitation events) that are not labelled as agitation. We also observe that there are many empty frames in the training set, where there are no people in the scene. Training the autoencoder on a large number of such frames could also lead to increased false alarm rate during testing. Future analyses can address this by under-sampling frames showing an empty hallway to avoid biasing the training of the autoencoder. Another approach could be, during training the autoencoder, adaptively re-labelling outliers from the IQR analysis as normal activities in order to avoid classifying them as anomalous events. This proof-of-concept analysis focuses on training the models using a single camera view. Future studies will explore training the models from multiple views; however, this may be computationally intensive. One way to address this may be to train the model on one camera view (e.g., hallway) and test on other camera view (e.g., dining hall), to make the overall anomaly detection task simpler and faster. From a modelling perspective, other advanced video anomaly detection algorithms [18] can also be evaluated, including long short-term memory and temporal convolutional networks to better capture the temporal information in the windows. Barriers to be addressed for future analyses include the size of the dataset (several terabytes of video) and the computational demands of the analysis. Graphics Processing Unit (GPU) clusters are needed for processing the surveillance videos, and running the computationally intensive 3DCAE model. In our case, Tesla P100 PCIe 12GB GPU clusters were used. The amount of time required to run the 3DCAE model was approximately 24 hours.

From a clinical perspective, this study is a first step towards the development of video-based clinical agitation detection systems with applications in clinical settings including long-term care, mental health inpatient and residential care. Agitation and aggression are clinically significant behaviours of risk in these environments, with a large impact in terms of workplace safety and adverse event prevention. The use of an anomaly detection framework is valuable in that it accommodates a range of different types of risky behaviour that might be observed in these environments, such as climbing on, moving or throwing furniture and banging on doors, in addition to physically aggressive behaviours directed towards others [38], [39]. It also does not require the identification of the individuals within the video stream. The downsides are that any novel or unusual visual stimuli will be triggered as events of interest, such as large pieces of equipment moving in the scene. Any clinical system based on this technology would need to have a way to handle anomalous ''alerts'' to minimize disruption from false positives.

The strength of this study is that it uses a well-annotated unique data set and a novel methodological approach. A limitation is that only a single camera view and participant were included. Another limitation is that the use of videos presents concerns around privacy and surveillance. Despite the widespread use of video surveillance in healthcare settings, this issue is far from settled [40], [41]. One possible approach to address this issue is to use privacy protecting vision modalities in future studies, such as depth, thermal, or infrared cameras. However, there are now examples of commercially available systems using videos to detect falls in use with demonstrated acceptability in residential care settings [42]. Further studies addressing approaches to the ethical design of video-based intelligent systems for dementia care are needed.

VI. CONCLUSION

Agitation as a behavioural symptom of dementia can be detected in video using anomaly detection in this proof-of-concept study. Future work will involve comparing with other competitive models and expand this analysis to a larger dataset with multiple camera views and participants to build more robust and generalizable anomalous behaviour detection system.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENT

The authors would like to thank Robin Shan, Program Services Manager, Specialized Dementia Unit, Toronto Rehabilitation Institute, in facilitating the study and providing with the necessary logistics support and also would like to thank the PwD and their families and the staff on the unit for taking part in the study.

REFERENCES

[1] World Health Organization (WHO). (2020). *Dementia*. Accessed: Jan. 19, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] H. C. Kales, L. N. Gitlin, and C. G. Lyketsos, "Assessment and management of behavioral and psychological symptoms of dementia," *BMJ*, vol. 350, p. h369, Mar. 2015.

[3] R. M. Keszycki, D. W. Fisher, and H. Dong, "The hyperactivity–impulsivity–irritiability–disinhibition–aggression–agitation domain in Alzheimer's disease: Current management and future directions," *Frontiers Pharmacol.*, vol. 10, p. 1109, Sep. 2019.

[4] J. Cerejeira, L. Lagarto, and E. Mukaetova-Ladinska, "Behavioral and psychological symptoms of dementia," *Frontiers Neurol.*, vol. 3, p. 73, May 2012.

[5] Canadian Institute for Health Information (CIHI). (2021). *Dementia in Long-Term Care*. Accessed: Jan. 20, 2021. [Online]. Available: https://www.cihi.ca/en/dementia-in-canada/dementia-care-across-the-health-system/dementia-in-long-term-care

[6] *Long-Term Care in Ontario: Fostering Systemic Neglect*, Ontario Council Hospital Unions, Toronto, ON, Canada, 2014.

[7] A. C. Cote, R. J. Phelps, N. S. Kabiri, J. S. Bhangu, and K. Thomas, "Evaluation of wearable technology in dementia: A systematic review and meta-analysis," *Frontiers Med.*, vol. 7, p. 1005, Jan. 2021.

[8] B. S. Husebo, H. L. Heintz, L. I. Berge, P. Owoyemi, A. T. Rahman, and I. V. Vahia, "Sensing technology to monitor behavioral and psychological symptoms and to assess treatment response in people with dementia. A systematic review," *Frontiers Pharmacol.*, vol. 10, p. 1699, Feb. 2020.

[9] S. S. Khan, S. Spasojevic, J. Nogas, B. Ye, A. Mihailidis, A. Iaboni, A. Wang, L. S. Martin, and K. Newman, "Agitation detection in people living with dementia using multimodal sensors," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 3588–3591.

[10] S. Spasojevic, J. Nogas, A. Iaboni, B. Ye, A. Mihailidis, A. Wang, S. J. Li, L. S. Martin, K. Newman, and S. S. Khan, "A pilot study to detect agitation in people living with dementia using multi-modal sensors," *J. Healthcare Informat. Res.*, vol. 5, no. 3, pp. 342–358, Sep. 2021.

[11] S. S. Khan, B. Ye, B. Taati, and A. Mihailidis, "Detecting agitation and aggression in people with dementia using sensors—A systematic review," *Alzheimer's Dementia*, vol. 14, no. 6, pp. 824–832, Jun. 2018.

[12] C. Berridge, J. Halpern, and K. Levy, "Cameras on beds: The ethics of surveillance in nursing home rooms," *AJOB Empirical Bioethics*, vol. 10, no. 1, pp. 55–62, Jan. 2019.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[14] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowl. Eng. Rev.*, vol. 29, no. 3, pp. 345–374, Jan. 2014.

[15] M. Brezovan and C. Badica, "A review on vision surveillance techniques in smart home environments," in *Proc. 19th Int. Conf. Control Syst. Comput. Sci.*, May 2013, pp. 471–478.

[16] J. Ma, Y. Dai, and K. Hirota, "A survey of video-based crowd anomaly detection in dense scenes," *J. Adv. Comput. Intell. Intell. Informat.*, vol. 21, no. 2, pp. 235–246, Mar. 2017.

[17] C. Brax, L. Niklasson, and M. Smedberg, "Finding behavioural anomalies in public areas using video surveillance data," in *Proc. 11th Int. Conf. Inf. Fusion*, Jun./Jul. 2008, pp. 1–8.

[18] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, Feb. 2021, Art. no. 104078.

[19] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 25, 2020, doi: 10.1109/TPAMI.2020.3040591.

[20] J. Nogas, S. Khan, and A. Mihailidis, "Fall detection from thermal camera using convolutional LSTM autoencoder," in *Proc. 2nd Workshop Aging, Rehabil. Independ. Assist. Living, IJCAI Workshop*, 2018, pp. 1–5.

[21] J. Nogas, S. S. Khan, and A. Mihailidis, "DeepFall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders," *J. Healthcare Informat. Res.*, vol. 4, no. 1, pp. 50–70, Mar. 2020.

[22] M. Martinez, D. Ahmedt-Aristizabal, T. Vath, C. Fookes, A. Benz, and R. Stiefelhagen, "A vision-based system for breathing disorder identification: A deep learning perspective," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6529–6532.

[23] H. Sharma, R. Droste, P. Chatelain, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 987–990.

[24] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki, K. Asada, S. Kaneko, A. Sekizawa, and R. Hamamoto, "Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning," *Appl. Sci.*, vol. 11, no. 1, p. 371, Jan. 2021.

[25] D. Ahmedt-Aristizabal, C. Fookes, S. Denman, K. Nguyen, S. Sridharan, and S. Dionisio, "Aberrant epileptic seizure identification: A computer vision perspective," *Seizure*, vol. 65, pp. 65–71, Feb. 2019.

[26] W. Wang, A. Tamhane, R. J. Rzasa, H. J. Clark, L. T. Canares, and M. Unberath, "Otoscopy video screening with deep anomaly detection," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 11597, M. A. Mazurowski and K. Drukker, Eds. Bellingham, WA, USA: SPIE, 2021, pp. 339–344.

[27] V. F. S. Fook, P. V. Thang, T. M. Htwe, Q. Qiang, A. A. P. Wai, M. Jayachandran, J. Biswas, and P. Yap, "Automated recognition of complex agitation behavior of dementia patients using video camera," in *Proc. 9th Int. Conf. e-Health Netw., Appl. Services*, Jun. 2007, pp. 68–73.

[28] B. Chikhaoui, B. Ye, and A. Mihailidis, "Feature-level combination of skeleton joints and body parts for accurate aggressive and agitated behavior recognition," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 6, pp. 957–976, Nov. 2017.

[29] S. S. Khan, T. Zhu, B. Ye, A. Mihailidis, A. Iaboni, K. Newman, A. H. Wang, and L. S. Martin, "DAAD: A framework for detecting agitation and aggression in people living with dementia using a novel multi-modal sensor network," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 703–710.

[30] J. Cummings, J. Mintzer, H. Brodaty, M. Sano, S. Banerjee, D. P. Devanand, S. Gauthier, R. Howard, K. Lanctôt, C. G. Lyketsos, E. Peskind, A. P. Porsteinsson, E. Reich, C. Sampaio, D. Steffens, M. Wortmann, and K. Zhong, "Agitation in cognitive disorders: International psychogeriatric association provisional consensus clinical and research definition," *Int. Psychogeriatrics*, vol. 27, no. 1, pp. 7–17, Jan. 2015.

[31] S. S. Khan and B. Taati, "Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders," *Expert Syst. Appl.*, vol. 87, pp. 280–290, Nov. 2017.

[32] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941.

[33] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[35] W. Che and S. Peng, "3D dual path networks and multi-scale feature fusion for human motion recognition," in *Proc. IEEE 4th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, vol. 1, Dec. 2019, pp. 698–704.

[36] W. Falcon. (2019). *Pytorch Lightning*. [Online]. Available: https://github.com/PyTorchLightning/pytorch-lightning

[37] S. S. Khan, M. E. Karg, D. Kulić, and J. Hoey, "Detecting falls with X-factor hidden Markov models," *Appl. Soft Comput.*, vol. 55, pp. 168–177, Jun. 2017.

[38] M. H. Woolford, S. J. Stacpoole, and L. Clinnick, "Resident-to-resident elder mistreatment in residential aged care services: A systematic review of event frequency, type, resident characteristics, and history," *J. Amer. Med. Directors Assoc.*, vol. 22, no. 8, pp. 1678–1691.e6, Aug. 2021.

[39] M. S. Lachs, T. Rosen, J. A. Teresi, J. P. Eimicke, M. Ramirez, S. Silver, and K. Pillemer, "Verbal and physical aggression directed at nursing home staff by residents," *J. Gen. Internal Med.*, vol. 28, no. 5, pp. 660–667, May 2013.

[40] A. Grigorovich and P. Kontos, "Towards responsible implementation of monitoring technologies in institutional care," *Gerontologist*, vol. 60, no. 7, pp. 1194–1201, Sep. 2020.

[41] A.-M. Dorsten, K. S. Sifford, A. Bharucha, L. P. Mecca, and H. Wactlar, "Ethical perspectives on emerging assistive technologies: Insights from focus groups with stakeholders in long-term care facilities," *J. Empirical Res. Human Res. Ethics*, vol. 4, no. 1, pp. 25–36, Mar. 2009.

[42] E. Bayen, J. Jacquemot, G. Netscher, P. Agrawal, L. T. Noyce, and A. Bayen, "Reduction in fall rate in dementia managed care through video incident review: Pilot study," *J. Med. Internet Res.*, vol. 19, no. 10, p. e339, Oct. 2017.

SHEHROZ S. KHAN received the B.Sc. degree in engineering and the master's and Ph.D. degrees in computer science, in 1997, 2010, and 2016, respectively. He is currently working as a Scientist with the KITE—Toronto Rehabilitation Institute (TRI), University Health Network, Canada. He is also cross appointed as an Assistant Professor with the Institute of Biomedical Engineering, University of Toronto (U of T). Previously, he worked as a Postdoctoral Researcher with U of T and TRI. Prior to joining academics, he worked in various scientific and researcher roles in the industry and Government jobs. He is an Associate Editor of the *Journal of Rehabilitation and Assistive Technologies*. He has organized four editions of the peer-reviewed workshop on AI in aging, rehabilitation, and intelligent assisted living held with top AI conferences (ICDM and IJCAI), from 2017 to 2021. His research is funded through several granting agencies in Canada and abroad, including NSERC, CIHR, AGEWELL, SSHRC, CABHI, AMS Healthcare, JP Bickell Foundation, United Arab Emirates University, and LG Electronics. He has published 49 peer-reviewed research papers and his research focus is the development of AI algorithms for solving aging related health problems.

PRATIK K. MISHRA received the master's degree in computer science and engineering from the Indian Institute of Technology (IIT) Indore, India, in 2020. He is currently pursuing the Ph.D. degree in biomedical engineering with the Institute of Biomedical Engineering, University of Toronto (U of T). Previously, he worked as a Research Volunteer with the Toronto Rehabilitation Institute, Canada, and as a Data Management Support Specialist at IBM India Private Ltd. He is currently working towards the application of machine learning for detecting behaviors of risk in patients suffering from dementia.

NIZWA JAVED received the B.S. and M.S. degrees in electrical engineering from the Institute of Space Technology, Pakistan, in 2017. She is currently pursuing the Ph.D. degree in electrical engineering and computer science with York University, Canada. She also worked as a Research Analyst with the Toronto Rehabilitation Institute, Canada, in 2020. Her research interests include developing robotic perception capable of carrying out complex tasks in the real world using sensory data and computer vision-based algorithms.

BING YE received the M.Sc. degree from Queen's University, Canada.

She has over ten years research experience in aging and technology. She has been working as the Research Manager with the Intelligent Assistive Technology and Systems Laboratory (IATSL), University of Toronto (U of T), since 2016. Her research interests include age-related diseases (i.e., dementia), quality of life of older adults, people with dementia and their caregivers, assistive technologies, and human-centred design. Specifically, she is interested in how technology can help people age gracefully and assist caregivers in supporting care and how technology can be developed/designed that reflect user's needs and requirements.

KRISTINE NEWMAN received the Bachelor of Science and Master of Science degrees in nursing from Queen's University, in 2003 and 2005, respectively, and the Ph.D. degree in nursing science from the University of Toronto, in 2012.

She received a Knowledge Translation Canada: Strategic Training Initiative in Health Research Post-doctoral Fellowship with McMaster University in health evidence, in 2013. She is currently an Associate Professor with the Faculty of Community Services, Daphne Cockwell School of Nursing, Ryerson University. Her program of research relates to knowledge translation—health evidence, gerotechnology, dementia awareness and intergenerational relations, young caregivers & their families, and formal & informal caregivers. She is an affiliated Researcher/a Professor with University Health Network—Toronto Rehabilitation Institute, and the Intelligent Assistive Technology and Systems Laboratory (IATSL), University of Toronto.

Dr. Newman is a Founding Member of the World Young Leaders of Dementia (WYLD) and leads the Spare a Thought for Dementia Collaboration (www.thoughtsfordementia.com).

ALEX MIHAILIDIS P.Eng., received the Ph.D. degree from the University of Strathclyde, Glasgow, U.K., in 2002. He is currently the Barbabra G. Stymest Research Chair of Rehabilitation Technology with the KITE Research Institute, University Health Network/University of Toronto (U of T). He is also the Scientific Director of the AGE-WELL Network of Centres of Excellence, which focuses on the development of new technologies and services for older adults. He is also a Professor with the Department of Occupational Science and Occupational Therapy and the Institute of Biomedical Engineering, U of T, where he holds a cross appointment with the Department of Computer Science.

He is very active in the rehabilitation engineering profession and is the Immediate Past President for the Rehabilitation Engineering and Assistive Technology Society for North America (RESNA) and was named a fellow of RESNA, in 2014, which is one of the highest honours within this field of research and practice. He is an Internationally Recognized Researcher in the field of technology and aging. He has published over 150 journals and conference papers in this field and co-edited two books: *Pervasive Computing in Healthcare* and *Technology and Aging*. His research interests include biomedical and biochemical engineering, computer science, geriatrics, and occupational therapy.

ANDREA IABONI received the D.Phil. degree from Oxford University, U.K., in 2002, the M.D. degree from the University of Toronto, Canada, in 2006, and completed the Residency in psychiatry from FRCPC, in 2011. She received a Fellowship and Sub-Specialty in geriatric psychiatry, in 2013. She is currently a Geriatric Psychiatrist and a Clinician-Scientist. She is also an Assistant Professor with the Department of Psychiatry, University of Toronto, and a Scientist with the KITE—Toronto Rehab Institute, University Health Network, Toronto, Canada. She is the medical lead of the Specialized Dementia Unit, Toronto Rehab.