A Sensor-Driven Visit Detection System in Older Adults' Homes: Towards Digital Late-Life Depression Marker Extraction

Narayan Schütz[†], Angela Botros[†], Sami Ben Hassen, Hugo Saner, Philipp Buluschek, Prabitha Urwyler, Bruno Pais, Valérie Santschi, Daniel Gatica-Perez, René M. Müri and Tobias Nef

***Abstract*—Modern sensor technology is increasingly used in older adults to not only provide additional safety but also to monitor health status, often by means of sensor derived digital measures or biomarkers. Social isolation is a known risk factor for late-life depression, and a potential component of social-isolation is the lack of home visits. Therefore, home visits may serve as a digital measure for social isolation and late-life depression. Late-life depression is a common mental and emotional disorder in the growing population of older adults. The disorder, if untreated, can significantly decrease quality of life and, amongst other effects, leads to increased mortality. Late-life depression often goes undiagnosed due to associated stigma and the incorrect assumption that it is a normal part of ageing. In this work, we propose a visit detection system that generalizes well to previously unseen apartments - which may differ largely in layout, sensor placement, and size from apartments found in the semi-annotated training dataset. We find that by using a self-training-based domain adaptation strategy, a robust system to extract home visit information can be built (ROC AUC=0.773). We further show that the resulting visit information correlates well with the common geriatric depression scale screening tool (ρ=-0.87, p=0.001), providing further support for the idea of utilizing the extracted information as a potential digital measure or even as a digital biomarker to monitor the risk of late-life depression.**

***Index Terms*—Telemonitoring, Pervasive Computing, Domain Adaptation, Self-Training, Late-Life Depression, Digital Biomarker**

I. Introduction

WITH a progressively ageing population in many countries, technology-supported ageing to promote independent living is becoming a topic of high economic and social interest [1]. One branch of research in this domain is concerned with pervasive computing based home monitoring [2]. Usually, this means the placement of unobtrusive, often contactless, ambient sensor systems in an older person's home, with the aim to improve home care and provide additional safety. Relevant literature suggests that older adults show good acceptance of contactless sensor systems - such as passive infrared (PIR) motion sensors [3]. Simple PIR motion as well as reed switch based door sensors have been widely used to unobtrusively monitor older adults [4]–[7]. And in comparison to cameras, audio or radar-based monitoring, they are only minimally privacy-invading.

Preliminary evidence even suggests better health outcomes for older adults with pervasive computing assistance, as opposed to a control group without the same [8]. Measures derived from pervasive computing systems may be used to monitor specific, health-relevant metrics, such as for cognitive function [9], [10] or physical activity. This information in turn could allow for early detection of health changes and better risk and disease management, or allow one to monitor the effects of interventions. Such medically relevant digital measures, derived by means of modern information technology (usually outside the clinical environment), are increasingly being referred to as digital biomarkers [11]–[13].

One area where such objective markers might have significant potential is late-life depression, a common condition in older adults that significantly decreases quality of life [14] and is associated with a wide variety of negative health outcomes, including increased risk of mortality [15] or cardiovascular disease [16]. Late-life depression often goes undetected as it is associated with a certain stigma or is wrongfully mistaken for normal ageing [17]. A particular component and risk factor of late-life depression is loneliness and its more objective correlate, social isolation [18]–[20].

Detecting behaviors that are associated with late-life depression, could thus help in providing valuable information to primary care providers, indicating whether further clinical screening could be necessary. Indicators for social isolation, measurable by pervasive computing systems, could be time spent outside the home or frequency and duration of home visits - particularly for older adults living alone. In this context, the former has been shown to be associated with perceived loneliness [21]. However, with declining mobility, this source of social interaction may become increasingly difficult to

Narayan Schütz, Angela Botros, Sami Ben Hassen, Prabitha Urwyler and Tobias Nef are with the ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

Hugo Saner is with the ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland, and also with the Department of Cardiology, University Hospital Bern (Inselspital), University of Bern, Bern, Switzerland

Philipp Buluschek is with DomoSafety S.A., Lausanne, Switzerland

Daniel Gatica-Perez is with the Idiap Research Institute, Martigny, Switzerland, and also with the School of Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Bruno Pais and Valérie Santschi are with the La Source, School of Nursing Sciences, HES-SO University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland

René M. Müri is with the Department of Neurology, University Neurorehabilitation Unit, University Hospital Bern (Inselspital), University of Bern, Bern, Switzerland, and also with the ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

[†] contributed equally to this work.

obtain. Home visits likely constitute another important form of social interaction for older adults, one that is not limited to mobility constraints. As a result, automatically identifying home visits as an objective measure of social isolation could be an interesting way to recognize community-dwelling older adults at risk of developing or already facing late-life depression.

In earlier work, we have shown that visit detection based on unobtrusive contactless sensors is feasible in community-dwelling older adults [22]. However, the results were obtained for only a very small number of participants and visits. In addition, the employed one-class support vector machine (OCSVM) [23] approach, trained on a single apartment, showed difficulties in generalizing to previously unseen apartments. Furthermore, the relationship between home visits and health-relevant indicators was not analyzed.

In this work, we aim to develop a robust visit detection system that adapts better to previously unseen apartments with different layouts and sensor placements. Moreover, we aim to evaluate the possibility of using the detected visit information as a potential digital biomarker for social isolation to better assess risk of late-life depression. Towards this end, we compare multiple learning strategies and employ concepts from semi-supervised learning [24] and domain adaptation [25]. The performance of the proposed approach is finally evaluated based on a real-world visit dataset including more than 20’000 hours of streaming data and 2’106 annotated nurse visits.

Our main contributions are:

1) Introduction of a sensor driven, unobtrusive visit detection system using a self-training based domain adaptation algorithm that can adapt to heterogeneous feature spaces.
2) Extensive evaluation of system performance across various approaches based on real-world data from free-living older adults.
3) Demonstration of the potential medical utility of the proposed system through evaluating not only visit detection performance but also correlations with medically relevant geriatric depression scale (GDS) values.

II. Materials and Methods

A. Data

The data used in this work stems from a home monitoring study, where modern pervasive computing technologies for use in older community-dwelling adults were evaluated. In total, 21 older adults were included for a target monitoring duration of 12 months (due to attrition, only 13 participants successfully finished the study). The research was conducted between January 2017 and July 2018. Participant recruitment aimed at representing a naturalistic sample of old, community-dwelling, and alone-living population in Switzerland. In Figure 1, a detailed participant recruitment flowchart is shown. The study was conducted based on principles of the Declaration of Helsinki and was approved by the responsible ethics committee of the canton of Vaud (CER-VD: Cantonal Ethics Committee of Vaud on Research involving humans; ID: 2016-00762). All participants signed and handed in their informed consent prior to study participation.

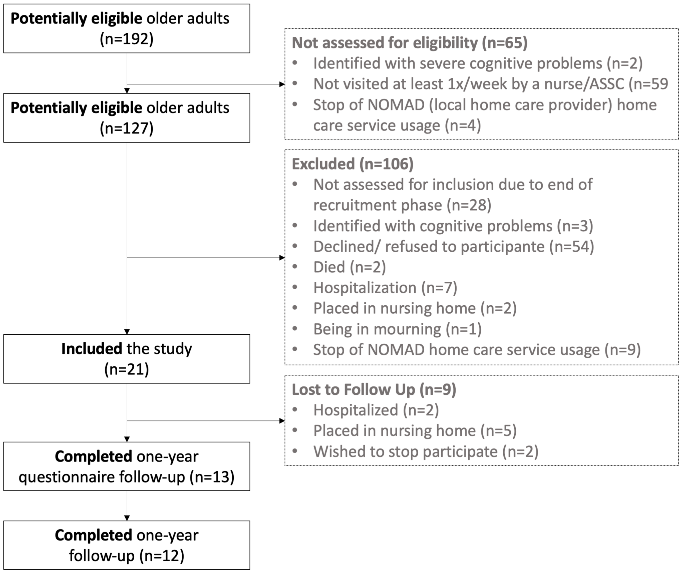
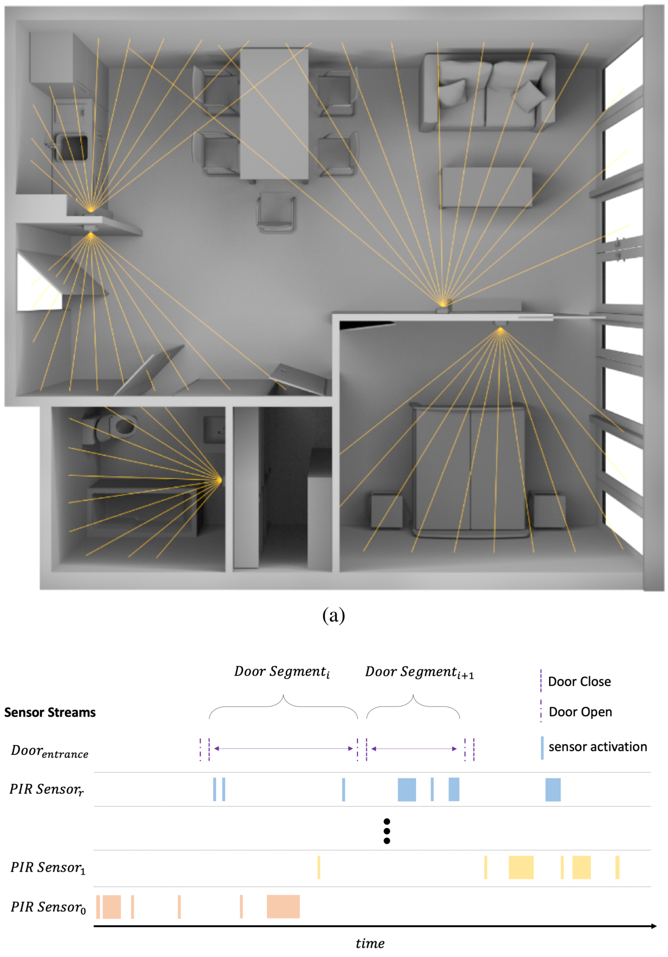

Potentially eligible older adults (n=192)
Not assessed for eligibility (n=65)
• Identified with severe cognitive problems (n=2)
• Not visited at least 1x/week by a nurse/ASSC (n=59
• Stop of NOMAD (local home care provider) home care service usage (n=4)
Potentially eligible older adults (n=127)
Excluded (n=106)
• Not assessed for inclusion due to end of recruitment phase (n=28)
• Identified with cognitive problems (n=3)
• Declined/ refused to participante (n=54)
• Died (n=2)
• Hospitalization (n=7)
• Placed in nursing home (n=2)
• Being in mourning (n=1)
• Stop of NOMAD home care service usage (n=9)
Included the study (n=21)
Lost to Follow Up (n=9)
• Hospitalized (n=2)
• Placed in nursing home (n=5)
• Wished to stop participate (n=2)
Completed one-year questionnaire follow-up (n=13)
Completed one-year follow-up (n=12)

Fig. 1. Displays the participant enrollment flowchart of the study within the used data has been collected [27].

Paticipant homes spanned small and medium-sized apartments and houses. Since we wanted to extract home visit information, the ground-truth for visits stemmed from nursing reports, marking when a NOMAD (Neuchâtel public home care association) nurse visited a participant. For this analysis all participants with labeled visits were included, leaving us with a set $\mathcal{Q} = \{q_1, ..., q_{15}\}$ of participants (age = 86 ± 7.23, sex = 54% female) and a total of 16’738 valid segments (equalling a total of 21’602 person-hours worth of sensor data - post pre-processing), of which 2’106 were annotated nurse-visit segments. The motion data was collected with a pervasive computing home-monitoring system (DomoSafety S.A., Lausanne, Switzerland), that comprises multiple PIR motion sensors that are placed in relevant rooms, as well as a magnetic entrance door and a fridge sensor. The PIR sensors report motion with a 0.5 Hz sampling rate, while the door sensors report every opening and closing event. If possible, PIR sensors were placed at a height of about 1.9m above the floor on a wall, in a way that their field of view would not detect motion outside a given room. For very large rooms, multiple sensors were mounted and assigned to the same room. All sensors communicate through the ZigBee [26] protocol with a base unit, which sends the data in real-time to the cloud via the cellular network. A schematic illustration of an installation in an apartment, as well as associated sensor streams, can be seen in Figure 2.

To extract features that could be used for visit detection, we first divide the monitoring time of a given participant q into a set of N time intervals, $\mathcal{T} = \{T_1, \dots, T_N\}$, based on entrance door opening and closing events, referred to as (door) segments, as illustrated in Figure 2. Subsequently, a *visit label* $y_i \in \{0, 1\}$ is assigned to each such segment T_i. Here, $y = 1$



(a)

(b)

Fig. 2. (a) gives an overview of the kind of sensor system used in an apartment (the apartment Layout has been adopted from [9]). (b) shows the resulting sensor streams that are being used throughout this work.

indicates a visit, and $y = 0$ no visit. This approach is based on the idea that the number of people in a home should not change without the opening and closing of the entrance door. Under the assumption that all home entrances are equipped with entrance door sensors, the resulting segments between two such opening-closing events could be treated as discrete entities with a single visit label [22]. Since the visit times reported by nurses were not perfectly accurate, we only label segments with at least 50% overlap between reported nurse visits and door segments as visit segments. All segments without labels are considered non-visits. It should be noted, that this leads to the inclusion of false labels, since there will also be non-nurse visits.

We further filter segments by their duration and time of occurrence, such that only segments with a duration between 5 minutes and 12 hours as well as occurrences between 4 am and 11 pm are included. This is done because we found it would lead to an over-estimation of system performance the other segments were included, as those segments are extremely unlikely to contain visits.

Then, each segment, T_i is assigned sequences of PIR sensor events, $E_i = (e_1, \dots, e_k, \dots, e_K)$, from sensors in the rooms $\mathcal{R}$. Note that $\mathcal{R}$ is the set of all hypothetically possible rooms, but apartments may only contain a subset of those. To construct shared features, $\mathfrak{F}$, we use a subset, $\mathcal{R}_\varepsilon \subset \mathcal{R}$, that we call *elementary* rooms. These include a living room, a toilet, an entrance, and a kitchen. These rooms are present across all apartments and thus domains. Individual PIR sensor events, $e_k \in \mathbb{N}^+$, describe the duration of the PIR activation time in seconds.

B. Self-Training based Domain Adaptation

To automatically detect visits, we are facing two major difficulties. First, we are dealing with considerable label noise, as we only have access to nurse-visits, while it is obvious that other visit types are also occurring and are arguably more important to detect. This could lead to a potentially significant number of wrongly labeled segments, depending on how many non-nurse visits a participant received. Additionally, this could also make a potential learning algorithm prone to specializing on nurse-visits, thus introducing what we further refer to as *nurse-bias*.

Second, it is difficult to calculate a set of general features to be used for visit detection which are comparable across different apartments - and individuals. Constituent factors here are different room compositions, altering apartment sizes, diverse sensor-placements, varying number of PIR sensors, and differing human behaviors. As a result, potential features have to be rather basic if the distribution between sources (installations with labels) and target (installation without labels) should be somewhat similar and thus useful for modelling, posing a certain risk of underfitting.

These two difficulties leave us with a complex heterogeneous feature space domain adaptation problem where no labels for the target domain are available and the source domain labels are potentially highly noisy. Formally, we have a set of multi-source domain samples, $\mathfrak{S} = \{(\boldsymbol{x}, y)_1^{q_1^{\mathfrak{s}}}, ..., (\boldsymbol{x}, y)_n^{q^{\mathfrak{s}}_{|Q \setminus \{q^{\mathfrak{t}}\}|}}\}$, from the $|Q \setminus \{q^{\mathfrak{t}}\}|$ participants in the training dataset, as well as a set of target domain samples from participant $q^{\mathfrak{t}}$, $\mathfrak{T} = \{\boldsymbol{x}_1^{q^{\mathfrak{t}}}, ..., \boldsymbol{x}_m^{q^{\mathfrak{t}}}\}$ that is treated as a new apartment in a leave-one-participant-out cross-validation procedure (details about the training procedure are presented in II-E). Here, $\boldsymbol{x} \in \mathcal{F}_q$ denotes a feature vector, m the number of samples in a target domain, and n the number of samples in the source domains. Additionally, feature vectors $\boldsymbol{x}$ from each source, as well as the respective target domain, stem from overlapping but heterogeneous feature spaces, $\mathcal{F}_q$ due to differing sensor and room combinations. As mentioned before, the OCSVM approach could not be generalized to some apartments, likely due to its inability to adapt to conditions of a specific target domain - a new apartment in our case.

To improve upon this, we adopt an approach similar in nature to an algorithm proposed in [28]. The concept is based on self-training [29] and was adapted to allow for domain adaptation. We further refer to this method as *ST-DA*. With ST-DA, we first train a simple base model $\mathcal{M}_{outer}$ on the labelled data from all source domains (we will henceforth refer to this as the "outer model"). The idea here is to train a high-bias model with limited learning capacity on a shared subset, $\mathfrak{F} = \bigcap_{i=1}^{|Q|} \mathcal{F}_{q_i}$, of rather generally valid features, that can be calculated across all domains. By choosing a simple model with limited learning capacity, we aim to reduce the risk of overfitting to specficially nurse-visits and force the model to use features that generalize across domains instead of modelling each source domain individually. To further reduce the *nurse-bias* risk, we try not to introduce any features into $\mathfrak{F}$ that could be very specific to nurse-visits, such as visit duration or time of day. Subsequently, high-confidence predictions of $\mathcal{M}_{outer}$ are used to pseudo-label a subset of samples from the target domain, $\mathfrak{T}$.

High-confidence here refers to the probability of a sample belonging to the visit or non-visit class, thus $Pr(y = 1|\mathbf{x})$ and $Pr(y = 0|\mathbf{x})$, respectively. The subset of pseudo-labeled samples is then defined as a certain percentage of samples with highest $Pr(y = 1|\mathbf{x})$ and highest $Pr(y = 0|\mathbf{x})$. The respective inclusion percentage, $\mathcal{P}_{selection} \in \{0.05, 0.15, 0.25\}$, was selected by means of the inner cross-validation loop (described in II-E).

These pseudo annotated samples are subsequently used to train a person and apartment specific inner model in a self-training fashion. This is achieved by iteratively adding high-confidence predictions to the pseudo annotated dataset and retraining the apartment specific model, $\mathcal{M}_{inner}$, on this new data. The resulting personalized model, $\mathcal{M}_{inner}$, (henceforth referred to as "inner-model") is eventually combined with the predictions from $\mathcal{M}_{outer}$ to get a final visit score for each segment. This final visit score is given by the segment duration weighted by the product of the inner and outer models' visit probabilities.

$$VS_i = w_i \cdot |T_i|$$
$$w_i = Pr_{\mathcal{M}_{inner}}(y_i = 1) \cdot Pr_{\mathcal{M}_{outer}}(y_i = 1)$$

where

- $Pr_{\mathcal{M}_{inner}}(y_i = 1)$ is the visit-scoring for a door segment, based on the inner model prediction
- $Pr_{\mathcal{M}_{outer}}(y_i = 1)$ is the visit-scoring for a door segment, based on the outer model prediction

One might interpret visit scores as weighted visit time. Note that we could have also calculated hard labels for visits. However, in this case we would loose information doing so, especially since a clear distinction between visit and non-visit may oftentimes not be possible with the incomplete information at hand.

While a product rule based combination of inner and outer model predictions should be more natural when combining probabilities, we additionally compare against the more commonly encountered mean aggregation rule. To test for statistical significance between the two approaches, we use an unpaired, two-sided, two-sample t-test.

For $\mathcal{M}_{outer}$ we evaluate two models. First an L1 penalized Logistic Regression (LR), because of its simplicity and associated limited learning capacity, as well as tendency to drive coefficients of unimportant features to zero, thereby potentially further increasing domain regularization. Second, a Local and Global Consistency (LGC) algorithm [30], also referred to as label spreading. LGC is a semi-supervised approach, similar to label propagation [31]. Broadly speaking, it iteratively assigns labels to data points based on a symmetrically normalized affinity matrix, the labels of neighbouring data points, as well as the initially given labels. This approach makes sense in our scenario since a certain amount of our non-visit instances are actually visits. Since we assume that visit samples generally lie on a certain, locally consistent, manifold, LGC can theoretically correct for the mislabeled data points, such as for instance shown in [32]. Note that due to computational complexity of running a nested cross-validation loop, we resort to the use of k-nearest neighbour based affinity matrices for LGC. Since $\mathfrak{F}$ is of limited dimensionality anyways, this should not realistically have a larger impact on results.

Unlike LR, which makes use of L1 regularization for feature selection, LGC has no inherent mechanism to perform feature selection. To still allow for feature selection, we include the same L1 penalized LR as a separate embedded feature selection step, when using the LGC algorithm. In this case, the LR used for feature selection is first optimized by means of the inner cross-validation loop and subsequently the set of features corresponding to non-zero coefficients is used for LGC training. Different combinations of ST-DA are reported as ST-DA[$\mathcal{M}_{outer}$, $\mathcal{M}_{inner}$].

For $\mathcal{M}_{inner}$, we evaluate multiple higher learning capacity models. First, a random forest (RF) classifier due to its generally good performance on a wide variety of learning tasks [33]. Second, a support vector machine (SVM) with Platt scaling based probability estimates, as it allows to implicitly solve the classification problem in a high dimensional space by means of kernel functions, where the two classes may be linearly separable. Third, a L1 penalized logistic regression with second order interaction terms (I-LR) due to its simplicity while still being able to potentially provide higher representational power as a result of the added interaction terms. A full description of the ST-DA Algorithm is given in Algorithm 1.

We hypothesize that the proposed approach has two major benefits: (1) the influence of label noise is reduced in the inner apartment specific model by using self-training and additionally due to the usage of LGC as the outer model. Furthermore, the bias towards nurse-visits is reduced by only including features that should be indifferent between nurse-visits and other visit types in the shared feature space $\mathfrak{F}$; (2) the higher learning-capacity inner model can adapt to individual apartment conditions, including its own feature space, $\mathcal{F}_q$, thereby leveraging additional information that is not used in $\mathcal{M}_{outer}$.

C. Shared Features in $\mathfrak{F}$

Here the features of the shared feature space $\mathfrak{F}$ are presented. $\mathfrak{F}$ is a subset of 12 features that are available across domains and potentially indicate the presence of additional

Algorithm 1: Outline Semi-Supervised Self-Training based Domain Adaptation

```
Initialize $\mathfrak{S}$, $\mathfrak{T}_{unlabelled}$, $\mathfrak{T}_{labelled}$, $\mathcal{M}_{outer}$, $\mathcal{M}_{inner}$,
  $\mathcal{P}_{selection}$;
$\mathfrak{S}$: labeled training dataset stemming from multiple
  sources (apartments with available nurse logs);
$\mathfrak{T}_{unlabelled}$: unlabeled dataset stemming from target
  distribution (new apartment);
$\mathfrak{T}_{labelled}$: dataset with self-trained pseudo labels for
  target dataset (new apartment);
$\mathcal{M}_{outer}$: outer model;
$\mathcal{M}_{inner}$: inner model;
$\mathcal{P}_{selection}$: high-confidence inclusion percentage;
Train $\mathcal{M}_{outer}$ classifier on labeled data $\mathfrak{S}$;
$\mathcal{M} \longleftarrow \mathcal{M}_{outer}$;
while $|\mathfrak{T}_{unlabelled}| \neq 0$ do
    foreach $(x_i, y_i) \in \mathfrak{T}_{unlabelled}$ do
        $p_i \longleftarrow P_{\mathcal{M}}(y_i = 1|\boldsymbol{x}_i)$;
        $y_i \longleftarrow f_{dichotomize}(p_i)$;
    end
    Sort $\mathfrak{T}_{unlabelled}$ based on prediction confidence $p_i$;
    Select a subset $\mathfrak{T}_{tmp} \subset \mathfrak{T}_{unlabelled}$ consisting of
      the $\mathcal{P}_{selection}$ percentage of the
      highest-confidence predictions;
    $\mathfrak{T}_{unlabelled} \longleftarrow \mathfrak{T}_{unlabelled} - \mathfrak{T}_{tmp}$;
    $\mathfrak{T}_{labelled} \longleftarrow \mathfrak{T}_{labelled} \cup \mathfrak{T}_{tmp}$;
    Train $\mathcal{M}_{inner}$ using $\mathfrak{T}_{labelled}$;
    $\mathcal{M} \longleftarrow \mathcal{M}_{inner}$
end
```

people in the apartment. What follows is an enumeration of those. Please note that while some features are motivated by probabilistic models, we do not necessarily emphasize theoretical correctness of the respective underlying assumptions.

1) $F_1^{\mathfrak{F}}$*: Normalized Location Sequence Log-Likelihood*
As visits likely lead to more uncommon room sequences, we try to capture this information through the normalized log-likelihood of a room transition sequence $S_i = ((r_j, r_k) \mid r_j, r_k \in \mathcal{R}_\varepsilon)$ given a maximum likelihood parameterized first order Markov chain (where parameters are estimated based on all available segments), with cardinality $|\mathcal{R}_\varepsilon| = R$. S_i represents the room transition sequence associated with segment T_i.
In essence this means that we calculate a zero diagonal transition matrix P based on the PIR sensor firing sequence and respective sensor localization.

$$P = \begin{pmatrix} Pr_{(r_1,r_1)} & Pr_{(r_1,r_2)} & \cdots & Pr_{(r_1,r_{R_\varepsilon})} \\ Pr_{(r_2,r_1)} & Pr_{(r_2,r_2)} & \cdots & Pr_{(r_2,r_{R_\varepsilon})} \\ \vdots & \vdots & \ddots & \vdots \\ Pr_{(r_{R_\varepsilon},r_1)} & Pr_{(r_{R_\varepsilon},r_2)} & \cdots & Pr_{(r_{R_\varepsilon},r_{R_\varepsilon})} \end{pmatrix}$$

where Pr_{r_j,r_k} represents the probability of transitioning from location (room) r_j to r_k, under the constraint that $Pr_{r_k,r_k} = 0, \forall\ k \ \in 1, ..., R$ and based on the assumption of first order Markov property.
Since the likelihood is dependent on the number of constituent factors, we use a length normalized variant, where $|S_i|$ is the number of transitions in segment T_i:

$$\frac{1}{|S_i|} \sum_{t_k \in S_i} log(Pr_{t_k}(t_k|t_{k-1}))$$

2) $F_2^{\mathfrak{F}}$*: Average Rare-Transition Probability*
We assume that visits lead to an increase in rare room transitions as it is more likely that PIR sensors form non-adjacent rooms will be triggered in short succession (e.g. when one person is in the bathroom, while another is in the kitchen and there is a living-room in between). Such transitions can not realistically be triggered by a single person and should thus be a good indicator for a visit. To capture this intuition, we assume the ten rarest transition counts to follow a Poisson distribution with rate parameter λ. This allows us to determine the maximum likelihood estimate of λ across all segments and then calculate the probability of a specific count in a segment.
Formally, we assume $F_2^{\mathfrak{F}} \sim \text{Pois}(\lambda)$. Here, $F_2^{\mathfrak{F}}$ stands for the random variable "rare room transition". The probability of observing a certain transition number h in a given time interval T_n is then defined by the respective probability mass function:

$$f(h, \lambda) = e^{-\lambda} \cdot \frac{\lambda^h}{h!}$$

with $\lambda = l \cdot |T_i|$ where :

- h is the number of times the sensors detected said transition within that specific segment.
- l is the rate parameter (number firings per second) calculated across all available segments of a person.

Subsequently we average the ten rarest transition count probabilities.

3) $F_3^{\mathfrak{F}}$*: Normalized Transition Duration*
Refers to the normalized time it takes to move from one room to another. Visits are thought to lead to faster transitions as it allows for 'teleportation' like behavior. Given an n-tuple D_i of room transition durations (can be thought of as the room sequence S_i but with associated transition times) in segment T_i, the normalized transition duration is calculated as follows:

$$\widetilde{D}_i = \frac{mean(D_i)}{\bar{D}}$$

where $\bar{D} \ = \ median((mean(D_i))_{i=1}^{N})$ is the normalizing factor. This feature is based on transitions between all available sensor equipped rooms $r \in \mathcal{R}$ of an apartment.

4) $F_4^{\mathfrak{F}}$ - $F_8^{\mathfrak{F}}$*: Normalized Activity Probabilities*
Describes the probability of obtaining a room activity value that is greater than the activity value of a given segment T_i. We observed that visits tend to lead to unusually high activity in certain rooms. This notion is calculated based on the empirical cumulative density

function (ECDF) of a users normalized activity $\hat{A}_q^r$ over all segments:

$$\hat{A}_q^r = \frac{A_q^r}{\bar{m}^r} \ \forall \ r \in \mathcal{R}_\varepsilon$$

where $\bar{m}^r = median(\{A_1^r, A_2^r, ..., A_N^r\})$ refers to the median activity of a room over all segments.

The ECDF is finally constructed over all normalized activity values $\hat{A}_i^r$. Given the ECDF, the probability of obtaining a larger normalized activity value than what was observed for a specific segment is then calculated as

$$Pr(X > \hat{A}_q^r) = 1 - ECDF(\hat{A}_q^r).$$

In addition to individual r the value is also calculated for the combined activity of all rooms in $\mathcal{R}_\varepsilon$.

5) *$F_9^{\mathfrak{F}}$ - $F_{12}^{\mathfrak{F}}$: Activity Percentages*

This feature represents the activity share of each included room $r \in \mathcal{R}_\varepsilon$ of a given apartment, per segment T_i, as well as the same for all rooms combined. It is assumed that visit segments do exhibit a different activity distribution among the rooms (people tend to spend time in specific rooms if they have visitors), compared to following usual daily routines.

$$\widetilde{A_q^r} = \frac{A_q^r}{A_q} \ \forall \ r \in \mathcal{R}_\varepsilon$$

where:

- $A_q^r = \frac{\sum_{e \in E_i} e}{|T_i|}$ is the duration of sensor activation per location r over the length of segment T_i.
- $A_q = \sum_r A_q^r$ is the total duration (over all rooms) of sensor activation in segment T_i.

D. Specific Features

Every apartment has a set of unique features in addition to the shared general ones. These are apartment specific or bear the potential to introduce a bias towards nurse-visits. This includes segment duration, hour of day, all of the shared features for additional rooms (where applicable), individual rare transition probabilities, and room transition durations for each transition type. The exact number of features thus varies from apartment to apartment.

E. Training and Evaluation

To evaluate the proposed approach with LR and LGC $\mathcal{M}_{outer}$, we compare against the OCSVM approach and in addition to cases of having only an outer model LR, LGC, RF, I-LR and SVM. With the OCSVM we use the normalized distance from the decision boundary as a way to quantify the likelihood of a segment belonging to the visit (or inlier) class. Outer model only classifiers are trained on the shared features in $\mathfrak{F}$. To assess performance of the different approaches in an unbiased manner and find respective hyperparameters, we employ a nested leave-one-person-out cross-validation loop. This means that for each iteration one participant is set aside as a test set, while training and hyperparameter optimization are performed on the remaining participants. To obtain realistic hyperparameter values, the same procedure is repeated within the training set. All features are normalized to have zero mean and unit variance. This is done on the basis of the respective training set. A detailed illustration is given in Figure 3. For training of both inner and outer models we used the respective scikit-learn implementations for LR/I-LR, LGC, RF, SVM and OCSVM (version: 0.23.1) [34].

Hyperparameters were obtained using grid-search. Table A1 shows the hyperparameter search space. Hyperparameters not mentioned were left at respective default values of the scikit-learn implementations.

As performance measure, with respect to detecting nurse-visits, we use the area under the receiver operating characteristic curve (ROC AUC) [35], since we are eventually interested in the system's ability to assign segments where $y = 1$ higher visit scorings than segments where $y = 0$. ROC AUC values are calculated directly on the basis of model predictions. In the case of ST-DA, this refers to the inner model. It should be noted here that a perfect ROC AUC value would indicate significant overfitting towards the nurse-visit sub-type and is therefore not desired. A good system should perform well with regards to ROC AUC values but not too well.

While we can not directly evaluate performance with respect to non-nurse visits, we aim to get an idea thereof by introducing the correlation with the GDS as a real-world performance measure. The GDS is a commonly used and well validated instrument to screen for late-life depression. The short version of the GDS results in a score between 0 and 15, where values above 5 are suggestive of depression. Introducing this additional metric is motivated by the idea that visits constitute an important part of social interaction and may therefore counteract isolation in older adults. As such, one could assume that participants with more overall visits, beyond just nurse visits, should be less likely to develop late life depression. As a result, we make the assumption that a visit detection system, which performs well with regards to general visit detection (not only on the nurse subset), should also exhibit the highest association with GDS scores. Moreover, our eventual goal for such a system is for it to be used to help in detecting social isolation and associated late life depression, not nurse-visits. To calculate this correlation, we use the partial non-parametric Spearman's rank correlation coefficient, controlling for the effect of potential confounding variables age, sex and mean nurse-visit duration. Since the GDS was a point-in-time measure we calculate the correlation with it based on the median of the daily visit scores DVS. The DVS are the sum of all individual segment scores VS_i of a given day. Partial correlations with GDS were calculated using R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) with package "ppcor" (version 1.1) [36]. Note that when the GDS is involved, we can only include participants that finished the one-year questionnaire followup, leaving us with $n = 13$ participants.

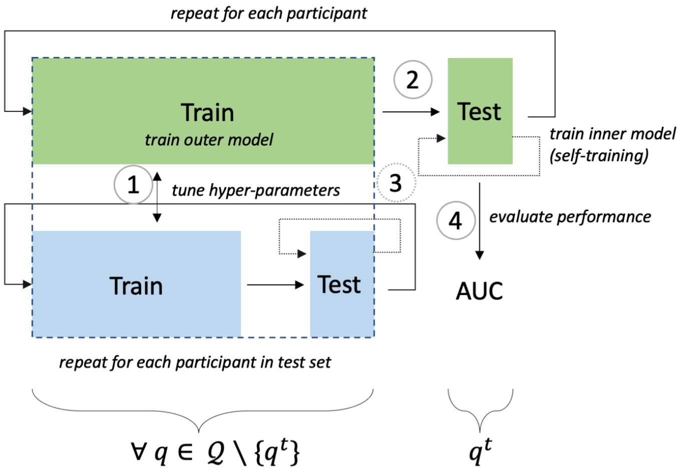


Fig. 3. Depicts the training procedure: (**1**) training dataset is subject to an inner leave-one-person-out cross-validation loop to find optimal hyperparameters to eventually train the outer model on; (**2**) the outer model is used to create initial high-confidence predictions to initially label data for the inner self-trained model of the left out (test-set) apartment; (**3**) the inner (apartment specific) model is trained using the semi-supervised self-training procedure; (**4**) visit scoring is applied and resulting predictions of the inner model are evaluated against real labels, resulting in a final ROC AUC value. The whole procedure is repeated for each apartment.

TABLE I
VISIT SCORE AGGREGATION COMPARISON

Approach	ρ Mean Rule	ρ Product Rule
ST-DA[LR, SVM]	-0.39	-0.49
ST-DA[LGC, SVM]	-0.31	-0.75
ST-DA[LR, I-LR]	-0.45	-0.44
ST-DA[LGC, I-LR]	-0.23	-0.82
ST-DA[LR, RF]	-0.32	-0.73
ST-DA[LGC, RF]	-0.41	-0.87
Mean	**-0.35**	**-0.68**

TABLE II
VISIT DETECTION PERFORMANCE

Approach	ROC AUC	Correlation GDS [ρ]
OCSVM	0.748	0.02 (p=0.952)
LR	0.736	-0.32 (p=0.375)
LGC	0.740	-0.17 (p=0.632)
RF	0.745	-0.45 (p=0.196)
SVM	0.747	-0.29 (p=0.412)
I-LR	0.739	-0.19 (p=0.653)
ST-DA[LR, SVM]	0.754	-0.49 (p=0.154)
ST-DA[LGC, SVM]	0.748	-0.75 (p=0.012)
ST-DA[LR, I-LR]	0.734	-0.44 (p=0.208)
ST-DA[LGC, I-LR]	0.744	-0.82 (p=0.004)
ST-DA[LR, RF]	0.774	-0.73 (p=0.023)
ST-DA[LGC, RF]	0.773	-0.87 (p=0.001)

III. RESULTS

Here we present results of the different approaches, compared to the proposed ST-DA approach, and amongst different variations of ST-DA. ROC AUC values represent the performance on the nurse-visit detection task, while associations with GDS values represent a proxy for more general real-world visit detection performance.

First, looking at the two aggregation rules for combining inner and outer model prediction to calculate visit scores (see Table I), it is visible that the product rule (average ρ=-0.68) leads to significantly higher (p=0.002) average partial correlations with the GDS, compared to the mean rule (average ρ=-0.35). In Table II, the AUC and partial GDS correlations of all approaches are given. Note that the respective correlations are based on the product rule, as it clearly performs better. In terms of the ROC AUC metric, ST-DA clearly exhibits the highest values of 0.774 and 0.773 for the LR and LGC variant with RF inner models, respectively.

With regards to the remaining combinations, ST-DA variants are showing, in most cases, higher ROC AUC values, when compared to the respective baseline approaches. However, the choice of inner model seems to matter here, as visible by the difference in ROC AUC values between the worst performing ST-DA[LR, I-LR] (ROC AUC=0.734) and the best performing ST-DA[LGC, RF] (ROC AUC=0.774)approach. The differences become more pronounced when looking at the partial correlations with the GDS values. Here, the baseline approaches are consistently worse, with the highest partial correlation of ρ=-0.45 (p=0.196) for the sole RF model and the lowest one for the OCSVM (ρ=0.02, p=0.952). The proposed ST-DA variants, showcase distinctly higher partial correlations, with ρ=-0.87 (p=0.001) being the highest value, achieved by the [LGC, RF] combination, and ρ=-0.44 (p=0.208) being the lowest, achieved by the [LR, I-LR] combination. Coming to the choice of the outer model, LGC based ST-DA variants have shown consistently higher partial correlations with GDS values, compared to LR variants. This is consistent across all ST-DA variants.

Figure 4 shows the association of visit scores with GDS values in form of a scatter plot of medians of daily visit scores, plotted against GDS values. Here the best performing ST-DA[LGC, RF] variant was used to derive the visit scores. It is well visible how higher visit-scores are correlated with lower GDS values.

IV. DISCUSSION

Based on the hypothesis that home visits can be an indicator for social isolation and associated geriatric depression, we introduce ST-DA, a self-training based domain adaptation approach that is specifically tailored towards the scenario of having only nurse-visit labels available for training a visit-detection system in a multi-source domain adaptation scenario with heterogeneous feature spaces. Using the OCSVM approach of previous work as a baseline, we compared against other two-class machine learning methods as well as multiple variants of the ST-DA approach. Furthermore, we evaluated performance on the nurse-visit detection labels based on ROC AUC values, as well as on overall partial correlations with GDS values, which serve as a proxy for real-world performance.

In terms of ROC AUC values, we found the existing OCSVM baseline to be in-line with the performance of binary-classifiers. However, there seems to be a large discrepancy when it comes to partial correlations with GDS

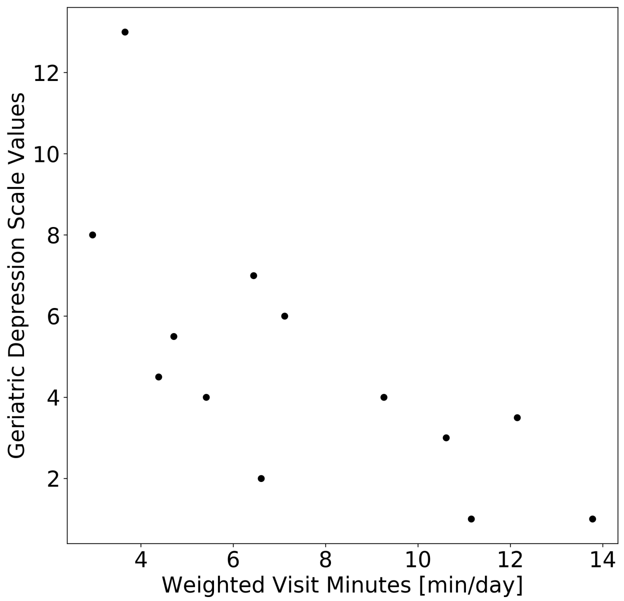


Fig. 4. Shows a scatter plot of the median of daily visit scores (weighted visits) against values from the geriatric depression scale assessments. Each dot represents one participant.

values. There, the OCSVM baseline performed drastically worse, showcasing zero association with GDS values. In this case, it appears that the benefit of additional information added by the inclusion of negative examples outweighs the introduction of additional label noise. We assume a major benefit of the two-class models is based on fact that the one-class approach tries to find a decision boundary around the nurse-visit class but the distance from this boundary does not necessarily imply that we simultaneously move closer to the non-visit class in some hyper-dimensional space.

Compared to both OCSVM and the two-class approaches, performance measures for the ST-DA variants are clearly higher. This supports the assumption that a main drawback of both OCSVM and binary approaches is the variance-bias like trade-off between overfitting to specific apartments (leading to poor generalization performance on largely different apartments) and relying on rather general features that are present across apartments but may lead to underfitting. With the ST-DA approach the inner model has access to additional apartment specific features, as well as features that were excluded because they could have introduced a nurse-bias. This not only allows a $\mathcal{M}_{inner}$ model to better adapt to features in $\mathfrak{F}$ of a specific apartment and the behavior of its inhabitant, but also gives it more potential information that is available exclusively in a given apartment's unique feature space $\mathcal{F}_q$. An additional factor explaining the better performance of ST-DA could be related to the impact of label noise, which may be corrected for when using the LGC $\mathcal{M}_{outer}$ model, and could also be further alleviated by the usage of pseudo-labels for $\mathcal{M}_{inner}$ training. While ST-DA with LGC and LR obtained very similar ROC AUC performance, the LGC variant achieved consistently higher partial GDS correlations. This supports the idea that LGC helps with label noise in the training data, which in turn increases real-world performance but therefore will not necessarily increase performance on the mislabeled nurse-visit sub-type.

When it comes to the combination of information from $\mathcal{M}_{outer}$ and $\mathcal{M}_{inner}$, the product rule leads to significantly higher partial correlations with GDS values. While we are not exactly sure why the difference is so clear, one reason may be the partially independent feature spaces, which could favor the product of probabilities, although this should only even play a factor in multi-class settings [37]. A practical and more probable factor may be that $\mathcal{M}_{inner}$ here acts as a gating mechanism. This behavior could thus, relatively speaking, help to build visit scores that are more pronounced and may in turn have higher discriminating power with respect to late-life depression.

Overall, in comparison to naïve baseline approaches, the higher ROC AUC as well as markedly higher GDS correlation performance of ST-DA suggest that by using self-training based domain adaptation it is possible to build a more robust visit detection system that adapts well to unseen apartments. Even more importantly, however, it demonstrates ST-DA's ability to extract late-life depression relevant information, in spite of significant label noise and varying source domains with heterogeneous feature spaces.

The extracted home visit information could eventually be used as a potential digital measure or even biomarker for primary care providers, indicating the level and evolution of loneliness and social isolation. This, in turn, could make it easier for healthcare professionals to assess the risk of an older adult to develop late life depression.

While our results are promising one should be aware that there are several limitations to this work. To further solidify the shown system and especially validate the potential of the calculated visit scores as a marker for late-life depression, larger datasets will be required. As such, while there is no obvious reason why the shown approach should not work for different population samples, results should be interpreted with caution and seen as a proof-of-concept. Additionally, a large part of our results is based on the assumption that the GDS is a good measure for late-life depression and that visits actually are associated with late-life depression. Furthermore, diagnosing depression was not the primary goal of the related study, no medical diagnosis has been performed on the subjects in this regard. While the GDS is a standardized and clinically well validated screening tool for late-life depression, it is a screening tool and no substitute for a complete medical assessment with a resulting diagnosis, conducted by a medical doctor.

V. Outlook

To further validate the shown approach, it will be necessary to have access to detailed medical diagnosis with regards to late-life depression, in addition to GDS based screening. Moreover, in order to quantify the effect of non-nurse visits in the results, beyond using GDS as a proxy measure, future work will have to test this methodology on data that contains ground-truth labels for these visits. This may be non-trivial and will require either active engagement of family members, or additional sensing (e.g. video outside the home), none of which are ideal approaches.

As a result, immediate next steps with regards to the proposed visit detection system will be to deploy the system in larger populations of older adults to define normative visit score values and define clinically meaningful cut-off values for social isolation and risk of depression in older adults. Here it will be highly important to specifically conduct medical diagnosis of late-life depression and obtain more accurate visit annotations - for at least a small subset of the population. Long-term it will also be of great interest to examine the temporal dynamics of visit scores over multiple years and potentially validate temporal changes against more objective biomarkers of depression, such as epigenetic, transcriptomic, proteomic or neuroimaging [38] based ones.

VI. Conclusion

We propose a home-visit detection system that adapts well to previously unseen apartments in a difficult multi-source domain adapation scenario with heterogenous feature spaces, nurse-visit sub-type bias as well as unquantified label noise. The underlying sensor system is conactless and based on unobtrusive passive infrared motion as well as door sensors, which protect the privacy of monitored subjects. Our results show that using a self-training based domain adaptation approach yields good performance with respect to visit detection, both in terms of ROC AUC values as well as with regard to task relevant real-world performance, corresponding to high and statistically significant partial correlations with geriatric depression scale values. This further indicates that the extracted visit information may indeed prove useful as a digital measuer or even biomarker for late-life depression, and that the visit detection system generalizes well - beyond the nurse-visits, on which it was trained. While we treated the case of visit detection, it is plausible that the proposed self-training based domain adaptation approach may be suitable for other pervasive computing scenarios where mislabeling and heterogeneous domains pose a challenge.

Acknowledgment

We would like to thank all participants and collaborators that made the collection of the used data possible. In particular, we would like to thank the NOMAD home care association and all involved nurses for their invaluable help in collecting the used data. Additionally, we would like to thank Michael Single for his valuable comments.

References

[1] S. Harper, "Economic and social implications of aging societies," *Science*, vol. 346, no. 6209, pp. 587–591, 2014.

[2] K. K. Peetoom, M. A. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, vol. 10, no. 4, pp. 271–294, 2015.

[3] C. S. Jacelon and A. Hanson, "Older adults' participation in the development of smart environments: An integrated review of the literature," *Geriatric Nursing*, vol. 34, no. 2, pp. 116–121, 2013.

[4] T. L. Hayes, F. Abendroth, A. Adami, M. Pavel, T. A. Zitzelberger, and J. A. Kaye, "Unobtrusive assessment of activity patterns associated with mild cognitive impairment," *Alzheimer's & Dementia*, vol. 4, no. 6, pp. 395–405, 2008.

[5] M. Popescu, B. Hotrabhavananda, M. Moore, and M. Skubic, "Vampir-an automatic fall detection system using a vertical pir sensor array," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 2012, pp. 163–166.

[6] S. M. Thielke, N. C. Mattek, T. L. Hayes, H. H. Dodge, A. R. Quiñones, D. Austin, J. Petersen, and J. A. Kaye, "Associations between observed in-home behaviors and self-reported low mood in community-dwelling older adults," *Journal of the American Geriatrics Society*, vol. 62, no. 4, pp. 685–689, 2014.

[7] M. J. Rantz, M. Skubic, S. J. Miller, C. Galambos, G. Alexander, J. Keller, and M. Popescu, "Sensor technology to support aging in place," *Journal of the American Medical Directors Association*, vol. 14, no. 6, pp. 386–391, 2013.

[8] M. J. Rantz, M. Skubic, M. Popescu, C. Galambos, R. J. Koopman, G. L. Alexander, L. J. Phillips, K. Musterman, J. Back, and S. J. Miller, "A new paradigm of technology-enabled 'vital signs' for early detection of health change for older adults," *Gerontology*, vol. 61, no. 3, pp. 281–290, 2015.

[9] N. Schütz, H. Saner, B. Rudin, A. Botros, B. Pais, V. Santschi, P. Buluschek, D. Gatica-Perez, P. Urwyler, L. Marchal-Crespo *et al.*, "Validity of pervasive computing based continuous physical activity assessment in community-dwelling old and oldest-old," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[10] P. Urwyler, R. Stucki, L. Rampa, R. Müri, U. P. Mosimann, and T. Nef, "Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living," *Scientific reports*, vol. 7, p. 42084, 2017.

[11] S. Meister, W. Deiters, and S. Becker, "Digital health and digital biomarkers–enabling value chains on health data," *Current Directions in Biomedical Engineering*, vol. 2, no. 1, pp. 577–581, 2016.

[12] A. Coravos, S. Khozin, and K. D. Mandl, "Developing and adopting safe and effective digital biomarkers to improve patient outcomes," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–5, 2019.

[13] A. Coravos, J. C. Goldsack, D. R. Karlin, C. Nebeker, E. Perakslis, N. Zimmerman, and M. K. Erb, "Digital medicine: A primer on measurement," *Digital Biomarkers*, vol. 3, no. 2, pp. 31–71, 2019.

[14] H. Sivertsen, G. H. Bjørkløf, K. Engedal, G. Selbæk, and A.-S. Helvik, "Depression and quality of life in older persons: a review," *Dementia and Geriatric Cognitive Disorders*, vol. 40, no. 5-6, pp. 311–339, 2015.

[15] E. Murphy, R. Smith, J. Lindesay, and J. Slattery, "Increased mortality rates in late-life depression," *The British Journal of Psychiatry*, vol. 152, no. 3, pp. 347–353, 1988.

[16] D. G. Blazer, "Depression in late life: review and commentary," *Focus*, vol. 7, no. 1, pp. 118–136, 2009.

[17] J. Unützer, "Late-life depression," *New England Journal of Medicine*, vol. 357, no. 22, pp. 2269–2276, 2007.

[18] J. Petersen, J. Kaye, P. G. Jacobs, A. Quinones, H. Dodge, A. Arnold, and S. Thielke, "Longitudinal relationship between loneliness and social isolation in older adults: Results from the cardiovascular health study," *Journal of Aging and Health*, vol. 28, no. 5, pp. 775–795, 2016.

[19] S. P. Greenstein, D. McGonigle, and C. H. Kellner, "Late-life depression," *Mount Sinai Expert Guides: Psychiatry*, pp. 312–321, 2016.

[20] R. Aylaz, Ü. Aktürk, B. Erci, H. Öztürk, and H. Aslan, "Relationship between depression and loneliness in elderly and examination of influential factors," *Archives of gerontology and geriatrics*, vol. 55, no. 3, pp. 548–554, 2012.

[21] J. Petersen, D. Austin, J. A. Kaye, M. Pavel, and T. L. Hayes, "Unobtrusive in-home detection of time spent out-of-home with applications to loneliness and physical activity," *IEEE journal of biomedical and health informatics*, vol. 18, no. 5, pp. 1590–1596, 2013.

[22] R. Hu, H. Pham, P. Buluschek, and D. Gatica-Perez, "Elderly people living alone: detecting home visits with ambient and wearable sensing," in *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*. ACM, 2017, pp. 85–88.

[23] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[24] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[25] H. Daume III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of artificial Intelligence research*, vol. 26, pp. 101–126, 2006.

[26] P. Kinney *et al.*, "Zigbee technology: Wireless control that simply works," in *Communications design conference*, vol. 2, 2003, pp. 1–7.

[27] B. Pais, P. Buluschek, G. DuPasquier, T. Nef, N. Schütz, H. Saner, D. Gatica-Perez, and V. Santschi, "Evaluation of 1-year in-home monitoring technology by home-dwelling older adults, family caregivers, and nurses," *Frontiers in public health*, vol. 8, 2020.

[28] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in neural information processing systems*, 2011, pp. 2456–2464.

[29] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models." *WACV/MOTION*, vol. 2, 2005.

[30] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[31] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Technical Report CMU-CALD-02–107, Carnegie Mellon University, Tech. Rep., 2002.

[32] B. K. de Aquino Afonso and L. Berton, "Analysis of label noise in graph-based semi-supervised learning," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 1127–1134.

[33] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014. [Online]. Available: http://jmlr.org/papers/v15/delgado14a.html

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[35] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[36] S. Kim, "ppcor: an r package for a fast calculation to semi-partial correlation coefficients," *Communications for statistical applications and methods*, vol. 22, no. 6, p. 665, 2015.

[37] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[38] R. Strawbridge, A. H. Young, and A. J. Cleare, "Biomarkers for depression: recent insights, current challenges and future prospects." *Neuropsychiatric disease and treatment*, 2017.

Appendix
Hyperparameter Search Space

TABLE A1

An overview of the optimized hyperparameters and the associated search space for each model used.*

Model	**Parameter**	**Values**
LR/ I-LR	C	[0.01, 0.1, ..., 10, 100]
LGC	n_neighbors	[1, 5, 10, 20, 50]
LGC	α	[0.05, 0.25, 0.5, 0.75, 0.95]
LGC	kernel	[knn]
OCSVM	γ	[0.01, 0.1, ..., 10, 100]
OCSVM	ν	[0.01, 0.25, 0.5, 0.75, 0.99]
OCSVM	kernel	[rbf, linear]
RF	max_depth	[None, 12, 6, 4]
RF	min_samples_split	[2,5,10]
RF	min_samples_leaf	[1,3,5,10]
RF	max_features	["auto", 5,10,20]
SVM	C	[0.01, 0.1, ..., 10, 100]
SVM	γ	[0.01, 0.1, 1, 10]
SVM	coef0	[0.01, 0.1, 1, 10]
SVM	kernel	[linear, polynomial, rbf]
SVM	degree	[2, 3, 4]

* for ease of reproducibility we are here adopting the nomenclature employed in respective scikit-learn implementations. Further note that LGC is called Label Spreading in the scikit-learn library.