An Activity Recognition System for Taking Medicine Using In-The-Wild Data to Promote Medication Adherence

Josh Cherian*
jcherian14@tamu.edu
Texas A&M University
College Station, Texas, USA

Samantha Ray*
sjr45@tamu.edu
Texas A&M University
College Station, Texas, USA

Tracy Hammond
hammond@tamu.edu
Texas A&M University
College Station, Texas, USA

ABSTRACT

Nearly half of people prescribed medication to treat chronic or short-term conditions do not take their medicine as prescribed. This leads to worse treatment outcomes, higher hospital admission rates, increased healthcare costs, and increased morbidity and mortality rates. While some instances of medication non-adherence are a result of problems with the treatment plan or barriers caused by the health care provider, many are instances caused by patient-related factors such as forgetting, running out of medication, and not understanding the required dosages. This presents a clear need for patient-centered systems that can reliably increase medication adherence. To that end, in this work we describe an activity recognition system capable of recognizing when individuals take medication in an unconstrained, real-world environment. Our methodology uses a modified version of the Bagging ensemble method to suit unbalanced data and a classifier trained on the prediction probabilities of the Bagging classifier to identify when individuals took medication during a full-day study. Using this methodology we are able to recognize when individuals took medication with an F-measure of 0.77. Our system is a first step towards developing personal health interfaces that are capable of providing personalized medication adherence interventions.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → **Machine learning algorithms**; • **Human-centered computing** → *Smartphones*; *Mobile devices*; **Mobile computing**.

KEYWORDS

human activity recognition, wearable technology, medication adherence, ADLs

ACM Reference Format:
Josh Cherian, Samantha Ray, and Tracy Hammond. 2021. An Activity Recognition System for Taking Medicine Using In-The-Wild Data to Promote Medication Adherence. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397481.3450673

*Both authors contributed equally to this research.

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450673

1 INTRODUCTION

One of the most common treatments for chronic conditions is the long-term use of medication. However, medication non-adherence remains a significant and widespread issue, as an estimated 50% of patients do not take their medication as prescribed [40]. These patients are subject to worsening clinical symptoms and outcomes, increased hospitalization readmission rates, and increased mortality rates [25, 39]. More specifically, medication non-adherence has been linked to a more than 2-fold increase in subsequent cardiovascular events in patients with coronary heart disease [20], increased hospitalization rates and total Medicare spending in patients with COPD [46], decreased glycemic control in patients with diabetes [10], and depression and lower quality of life in patients with Parkinson's disease [21]. Overall, medication adherence has become a significant strain on the healthcare industry, as an estimated $100 to $300 billion is spent annually in the United States on dealing with the consequences of this issue [25].

Given the prevalence and significance of medication non-adherence, researchers have proposed a variety of interventions in an effort to improve adherence rates. Many have proposed intervening through the use of electronic medication packaging or the integration of electronic devices into medical containers such as pill bottles, pillboxes, and blister packs [11]. One study found that the use of electronic prescriptions over paper prescriptions improved adherence [48], while studies that have allowed patients to track their blood pressure found mixed results when evaluating its effect on medication adherence [45]. One of the more common strategies in recent years has been the development of mobile applications [15, 36]. The most trustworthy apps according to pharmacists are Mango Health, MyMeds Medication Management, MediSafe Medication Management, and Dosecast Medication Reminder [49]. Most of these solutions are broad strokes in which opportunities for personalization remain limited, ultimately limiting their utility. To make these interfaces more convenient for their users and further advance over traditional methods, the next step is to make these interfaces intelligent and capable of automatically logging the users' behavior.

Human activity recognition (HAR), or the recognition of human activities through the analysis of wearable inertial or ambient sensor data, offers a means of facilitating personalized medication adherence interventions. Combining knowledge of the prescription with detection of how and when patients take their medication could feasibly power proactive health management interfaces that are able to provide timely, personalized interventions. This type

of interface could be integrated into daily routines in a helpful and non-disruptive manner, borrowing from and improving on the interfaces and user experiences of widely popular fitness trackers and smartwatches [15, 36]. In recent years, activity recognition researchers have made significant strides towards demonstrating that advanced activity recognition techniques can also be used to power such interfaces by recognizing and monitoring a wide variety of daily self-care activities (commonly referred to as Activities of Daily Living (ADLs)) [6, 14, 30]; in this work we aim to extend this body of work to include the accurate real-time recognition of the action of taking medication.

To achieve this we seek to address one of the larger outstanding research questions in the field of activity recognition applied to the activity of taking medication: Can we develop an activity recognition system capable of recognizing ADLs in an unconstrained, real-world environment? This task is challenging due to the inherent data skew of activities of interest to out-of-scope activities, e.g., the action of taking medication typically takes a few seconds, which is a tiny fraction of an entire day. Traditional methods for dealing with unbalanced classes result in suboptimal models in this domain as they generally boost the recall of the minority class at the cost of the classification precision. To address this we introduce portfolio classification, which uses the probabilities generated by a baseline classifier to improve performance, combined with a modified version of bagging in which we sample the classes independently to recognize the activity of taking medication in real-time.

2 RELATED WORK

2.1 Activity Recognition for Taking Medication

Activity recognition techniques have been used to recognize a wide variety of activities, most commonly leveraging wearable inertial sensors in the form of accelerometers, gyroscopes, and/or microphones and machine learning techniques to recognize ambulation activities [6, 29, 44], fitness activities [35, 37], and other ADLs [14, 30, 50]. There has been a considerable amount of attention paid specifically to taking medication by activity recognition researchers given the prevalence and consequences of medication non-adherence [40]. It's worth noting that most of these approaches have looked at the physical motion of taking medication, rather than the overarching ADL of taking medication (one exception is Dernbach et al. [16] where subjects retrieved and sorted out medication doses for a week). Many of these studies have used sensor data collected from smartphones and smartwatches [18, 23, 26, 27, 34, 52]. Kalantarian et al. in particular has explored the idea of using wearable technology from multiple perspectives, analyzing the different ways people take medications [27], using smartwatch-based detection [26] and using a wearable neck sensor [28]. Several studies have taken alternative approaches, developing systems not worn by the individual. A study by Aldeer et al. [3] placed accelerometers in the bottle and cap to differentiate users, while several studies have used the Microsoft Kinect to detect the motion of taking medication [12, 38]. Recent works have leveraged using distributed computing [18, 34] and have sought to integrate systems into a smart home environment [4].

2.2 Improving Classifier Performance

As classification tasks are one of the most common real-world applications of machine learning, a significant amount of emphasis has been placed on both developing better-performing algorithms and augmenting existing algorithms to reduce their error. Solutions focusing on the latter are diverse: even within the field of HAR, techniques range from rules based on domain knowledge, ensemble learning, and other semi-supervised techniques [1]. Using domain knowledge, researchers have determined the relevant features for HAR algorithms [27] or applied domain-specific rules to post-process the classified sensor data [14]. The use of domain knowledge is even more common in ambient sensing approaches to HAR [13].

Researchers have explored a number of different ensemble learning techniques. These include Co-training [22, 32, 47], a Genetic Algorithm-based Classifier Ensemble Optimization Method [17], a Hierarchical Weighted Classifier [5], Multi-view Stacking [19], and Voting [42, 43]. Other semi-supervised techniques include further training algorithms after deployment [33], and combining Multiple Eigenspaces with Support Vector Machines [24].

To our knowledge, no study has sought to recognize the activity of taking medication in an unconstrained, real-world environment using data collected from an in-the-wild user study. Furthermore, ensemble learning techniques have not been applied extensively to the activity of taking medication.

3 DESIGN MOTIVATION

Human activity recognition of ADLs is integral to developing applications for proactive health management. However, current research in this field generally focuses on the ability to distinguish a large number of activities only in the context of the activities within the dataset. In other words, there has been little focus on recognizing activities in in-the-wild settings where rejecting unknown activities is central to a system's success. This task is challenging due to the inherent data skew of activities of interest to out-of-scope activities. Traditional methods for dealing with unbalanced classes result in suboptimal models in this domain as they generally boost the recall of the minority class at the cost of the classification precision. In contrast to an unbalanced recognition task such as spam detection where false positives, i.e., false alarms, are preferable to false negatives, i.e., misses, a system that claims a user performed the "take medication" activity multiple times is worse than a system that misses when the user took medication. The former creates garbage data which results in additional problems for the user and their healthcare provider when reading the logs of the recognition system; the latter does not perform its intended purpose but at least does not make the problem worse. In scenarios where that minority class is not notably distinct from the majority class, e.g., the activity of taking medication when compared to many other activities performed throughout one's average day, this data skew becomes a serious barrier. Many samples of the class of interest will be ambiguous. Training a model on such data and expecting it to learn the minority class is asking the model to make a suboptimal choice. As such, training such a model on classification error will result in a model that favors the majority class.

A solution to combat this problem lies in using randomness to explore the data. In classification tasks, the two standout examples

of using randomness as a core component of their algorithm design are Random Forest [9] and Bagging [8]. Both of these methods train multiple classifiers and aggregate the results to create a classifier with lower variance and a lower chance of overfitting compared to using a single classifier. Random Forest trains multiple classifiers with all of the data but with random subsets of the features. Bagging, i.e., bootstrap aggregating, involves sampling the training set with replacement, i.e., bootstrapping, to create multiple classifiers on varied subsets of the training data. However, the standard Bagging technique relies on the assumption that each sampled subset will be reasonably representative of the overall data distribution. When applied directly to imbalanced datasets, Bagging will amplify the individual classifier's tendency to favor the majority class. To summarize, Random Forest uses randomness during feature selection for its internal classifiers, and Bagging uses randomness to select the training data of its internal classifiers.

As the taking medication action is difficult to distinguish from other activities performed throughout the day, Bagging has an edge over Random Forest with respect to creating internal models that can recognize the activity of taking medication. Taking samples of the data functions similarly to undersampling in addressing the data skew. But, Bagging does not work well with skewed data as it inherently favors the majority class. This favoritism of the majority class can be addressed with a simple modification: sample the classes independently so that the data imbalance can be controlled in each classifier within the aggregate Bagging classifier. As a result, each classifier is guaranteed to be familiar with the class of interest. Repeating this process k times results in a voting classifier where each classifier has a different perspective of the universe.

4 DATA COLLECTION

4.1 System Implementation

A major component of wearable health monitoring design involves the sensor type and placement. For a real-world system, the device must not inconvenience the user or cause discomfort. With this in mind, we selected smartwatches as the basis of our system to collect 3D accelerometer data in a unobtrusive fashion. Specifically, we used Polar M600 smartwatches with a custom-built data collection application that transmitted sensor data to Android smartphones in real-time. The data collection interface is shown in Figure 1. In addition to storing the data, the application was used to start and stop the data collection process and to label each activity performed. Participants could select from several pre-made buttons, e.g., brushing teeth, drinking, eating, clapping, taking medication, and washing hands, or write in a custom label, e.g., folding laundry. Because people can perform the activity of taking medication with both hands, the participants wore a watch on each wrist.

4.2 Activities

During the user study we asked the participant to label each activity they performed. If the participant did not want to label an activity, a default option of "nothing" was provided. We asked the participants to label five activities of interest: brushing teeth, drinking, eating, taking medication, and washing hands. These activities were selected because they are ADLs and involve specific movements of the hands and wrist. Brushing teeth and washing hands are

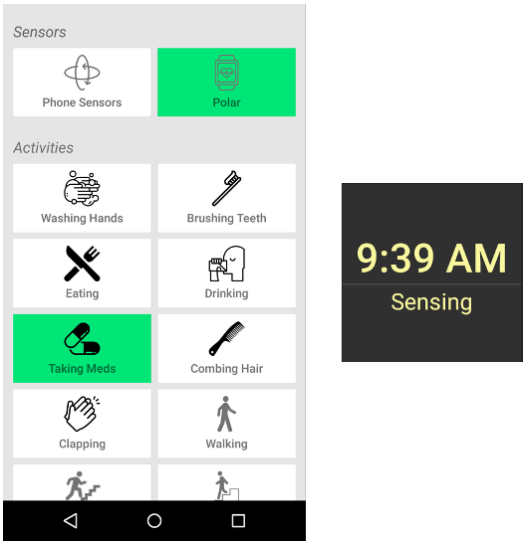

Figure 1: Smartphone and smartwatch interfaces for data collection. Users can start and stop data collection by selecting the sensors (in this study only data from the Polar smartwatches was used). When performing an activity users click the appropriate button to indicate what activity they are doing. When they finish performing the activity they could click the button again to indicate they had finished. Green icons indicate the button is currently selected. The watch interface simply indicates when data is being collected.

both personal hygiene ADLs that are comprised of rapid, repetitive movements. Conversely, eating, drinking, and taking medication are all Basic or Instrumental ADLs that consist of specific, non-repetitive movements.

Having these labels gives important information for recognizing the activity of taking medication. For one thing, the activities of eating and drinking are physically and contextually associated with the activity of taking medication. The atomic actions of bringing objects to one's mouth are performed when eating, drinking, or taking medication, and many medications require that people take them with water and/or food. The other main benefit of having additional labels than just the taking medication activity is having context for the participant's behavior during data analysis.

4.3 User Study

We recruited 9 participants (aged 18-30; 2 female, 7 male) for data collection. Participants were asked to wear the smartwatches and label their data for an entire day, i.e., from the time they got up in the morning until the time they went to bed at night. This type of study has the benefit of providing more realistic data; however, this comes at the cost of potential noncompliance and noisier data. For example, some participants completed the user study before the evening and returned the watches after wearing them for only part of their day. As shown in Table 1, the participants wore the watches for anywhere from 3 to 34 hours, resulting in approximately 94 hours of total collected sensor data.

On that note, participants were asked to integrate the activities of interest into their day if they did not already have the habit of doing them. This request does not make the data unnatural as all of the activities are ones that people should perform on a daily basis to stay clean and healthy. The only activity that is understandable that a person may not perform regularly is the main activity of interest, taking medication. As medication should not be taken for no reason, participants were provided with a prescription bottle filled with small candies, e.g., M&M's, to serve as a placebo medication for those who did not have any medications to take.

About 96% of the collected data was labelled "nothing" or "unknown," about 3% of the data was labeled a different ADL, and approximately 0.29% of the collected data was labelled "taking medication." Participants varied in the number of times they took medication throughout the study from 1 to 14 times; the amount of "taking medication" samples compared to the total number of samples from each participant varied from 0.08% to 0.91%. The distribution of activities across the participants is given in Table 1.

Table 1: Distribution of Activities Across Participants

Participant	Data Size (Hrs.)	% Taking Medication	# Times
P1	6.42	0.0008	1
P2	5.84	0.0035	1
P3	7.71	0.0034	2
P4	3.03	0.0066	2
P5	34.30	0.0014	5
P6	13.32	0.0030	6
P7	8.98	0.0016	2
P8	6.24	0.0039	6
P9	7.97	0.0091	14
Total	93.81	0.0028	39

5 BASELINE MODEL

To recognize when participants took medication we developed a baseline model following a traditional activity recognition approach. We extracted a total of 106 features from sliding windows of data, which were fed into a modified controlled Bagging classifier. Hyperparameter tuning was done using nested leave-one-subject-out cross-validation (LOSO CV) and the Tree-structured Parzen Estimator (TPE) algorithm which selects parameters for testing based on the performance of previously selected parameters. This algorithm often produces better results than other methods [7]. Based off the performance of various window sizes from 500ms to 5000ms and overlap percentages from 0% to 75% during hyperparameter tuning, the window size was set to 1000ms and the windows overlap by 75%. This window size is small enough to detect the activity of taking medication and the relatively high overlap percentage is useful for detecting atomic actions.

The classifier for the system used a cross-validation-based voting technique. After splitting the data into the training set and the evaluation set, the training set was split into k folds. Each of the k models were trained with a different fold held out as a development test set. The averages of the prediction probabilities for each class across the k models were used to determine the final prediction probabilities and the resulting prediction.

5.1 Features

The features we extracted from segmented accelerometer data are shown in Table 2. These features have either been used previously in activity recognition studies [14, 31] or were adopted from other domains such as audio analysis in the case of features (O)–(Q) (Eqs. 1 to 3) [51] and eye-tracking in the case of feature (S) (Eq. 4) [2]. Definitions for Features (A)–(N) and (R) match those used in prior activity recognition literature [14, 31]. Features (A)–(Q) were calculated for each axis of the accelerometer for each hand and features (R) and (S) were calculated for each hand, bringing the total number of features to 106 ((17 features $*$ 3 axes $*$ 2 hands) + (2 features $*$ 2 hands)).

In Eqs. 1 to 3, $X_i(k)$ represents the amplitude of the kth bin of the DFT spectrum. Feature (S) was estimated using the Minkowski-Bouligand box-counting method shown in Eq. 4 where D is the fractal dimension and $N(\epsilon)$ is the number of boxes of length ϵ required to cover the accelerometer path A.

$$C = \frac{\sum_{k=1}^{N} kX(k)}{\sum_{k=1}^{N} X(k)} \tag{1}$$

$$S = \sqrt{\frac{\sum_{k=1}^{N} (k-C)^2 X(k)}{\sum_{k=1}^{N} X(k)}} \tag{2}$$

$$\arg\min_{SR} \sum_{k=1}^{SR} X(k) \geq 0.85 \sum_{k} X(k) \tag{3}$$

$$D(A) := \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)} \tag{4}$$

5.2 Controlled Bagging

As described in Section 3, controlled Bagging is a modification on the Bagging ensemble machine learning model to enable recognition of rare classes. When training a controlled Bagging model, the training data is split into two subsets based on whether the samples belong to the minority class of interest or not. For each interior classifier, training data for each subset is sampled independently and then merged into a single training set to train the classifier as usual. If desired, the class of interest subset can be used in its entirety. The majority class is sampled based off the size of the class of interest subset multiplied by the algorithm parameter *skew*. *Skew* determines how rare the class of interest appears to each interior classifier; in other words, *skew* controls the undersampling rate of the majority class. The prediction probabilities of each interior classifier are aggregated to determine the Bagging classifier's prediction probabilities. Figure 2 depicts how the *skew* parameter affects performance on the test set within each fold during training. These F1-scores represent the performance of a single classifier on unseen data. Controlling the data skew between 2% and 10% results in improvements over the baseline of the majority classifier's performance, 0.5.

The controlled Bagging algorithm is written as a user-defined classifier in Python. Preprocessing steps and the base classifier leverage the scikit-learn packages [41]. We performed subset selection to reduce dimensionalality and improve generalization to the test set. Subset evaluation metrics tested include χ^2, classification F1 score, and mutual information. χ^2 ultimately produced the best

Table 2: Features Used in Primary Model

#	Feature Name		
(A)	Average Jerk	(K)	Standard Deviation of Number of Peaks
(B)	Average Height	(L)	Number of Valleys
(C)	Standard Deviation Height	(M)	Average Number of Valleys
(D)	Energy	(N)	Standard Deviation of Number of Valleys
(E)	Entropy	(O)	Spectral Centroid
(F)	Average Acceleration	(P)	Spectral Spread
(G)	Standard Deviation Acceleration	(Q)	Spectral Rolloff (<85%)
(H)	Root-Mean-Squared (RMS) Acceleration	(R)	Axis Overlap
(I)	Number of Peaks	(S)	Fractal Dimension
(J)	Average Number of Peaks		

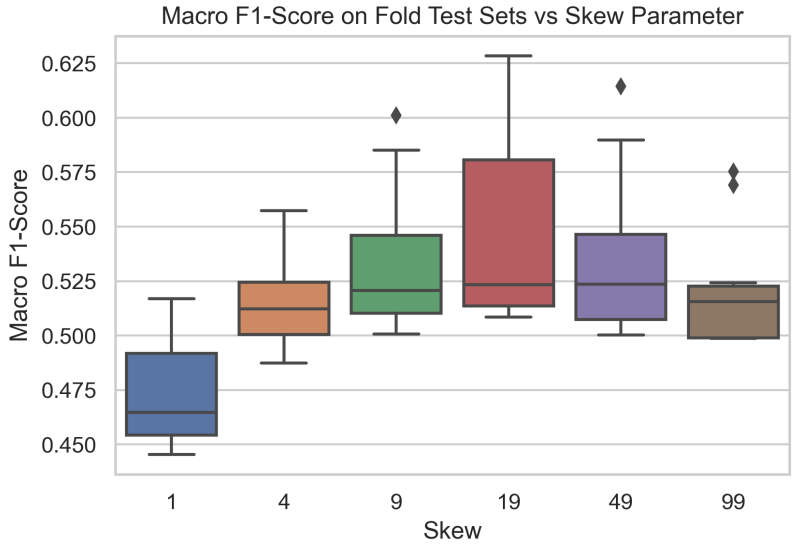

Figure 2: The impact of *skew* parameter on the classifier's performance on each fold's test set. These skew parameters represent a controlled data distribution for the "taking medication" class in the training data of 50%, 20%, 10%, 5%, 2%, and 1%.

performing models. To synergize with the χ^2 subset selection, we used min-max scaling as a preprocessing step the data for the model. Decision trees are used as the base classifier within the bagging classifier due to its simple and lightweight design. Each Bagging classifier uses 100 decision trees with the entropy criterion and a minimum of two samples per leaf to improve generalization.

5.3 Results

5.3.1 Baseline. To establish a baseline of performance more meaningful than the majority classifier's performance on this dataset, we trained a model only on a specific user's data for each of the 9 participants. The evaluation set for each of these models comprises of a stratified random 10% of that user's data, i.e., 10% of their "taking medication" samples and 10% of the majority "nothing" class. As shown in Table 3, the performance varies across the participants correlating with the percentage of that user's data being the "taking medication" activity.

Table 3: Baseline Performance: User-Specific Models

Participant	Macro F1-Score
P1	0.726
P2	0.565
P3	0.892
P4	0.705
P5	0.691
P6	0.745
P7	0.758
P8	0.685
P9	0.845
Average	0.735

5.3.2 User-Dependent Model. Our primary model is trained in a user-dependent fashion, i.e., data from each user appears in both the training and evaluation sets. The evaluation set for each of these models comprises of a stratified random 10% of each user's

data, i.e., 10% of their "taking medication" samples and 10% of the majority "nothing" class. This model does not improve over the established baseline with a macro-F1 score of 0.721 as shown in Table 4. Performance on each participant's data provided for the sake of comparison to the baseline; some user's individual performances improved while others worsened. This outcome implies that people perform the activity of "taking medication" differently, creating confusion within the multi-user model.

Table 4: User-Dependent Model

Participant	Macro F1-Score
P1	0.718
P2	0.604
P3	0.682
P4	0.602
P5	0.642
P6	0.666
P7	0.787
P8	0.687
P9	0.838
Overall	0.721

Table 5: User-Adaptive Model

Participant	Macro F1-Score
P1	0.595
P2	0.786
P3	0.695
P4	0.579
P5	0.636
P6	0.811
P7	0.602
P8	0.812
P9	0.730
Average	0.693

5.3.3 User-Adaptive Model. Because the "taking medication" classifiers did not generalize well to unseen users, we additionally trained classifiers that tune the user-dependent classifier to a specific user. As the Bagging classifier uses voting to determine its final prediction probabilities, combining the personalized recognition into the model is simple: a model trained only on one specific user also votes in the final classifier. As shown in Table 5, this model is worse than the sum of its parts.

6 SECONDARY MODEL

6.1 Portfolio Classification

In this work we introduce the concept of portfolios, on top of which we calculate simple heuristics to reduce the error produced by baseline classification algorithms. We define a portfolio as follows: given a set of activities $A = \{A^1, ..., A^N\}$ the baseline algorithm generates a set of probabilities or *portfolio* $P_{A'} = \{p_{A^1}, ..., p_{A^N}\}$ for each class A' where p_A is the average probability of predicting class A for class A' and $\sum_{i=1}^{N} p_{A^i} = 1$ for each portfolio. When used in k-fold cross-validation, portfolios for each class A' are unioned together to create a new portfolio of dimensionality $k * N$. In addition, the standard deviations for each portfolio $\sigma_{A'} = \{\sigma_{A^1}, ..., \sigma_{A^N}\}$ are collected to track the range of the prediction probabilities of each class on each class. In this work we look at a set of activities that are generally associated with similar contexts and movements of the hands and wrist, as these are the activities that systems would have trouble distinguishing in real world scenarios. We base our analysis on the intuition that not only will the portfolios of these classes feature high probabilities for the other gesturally-similar classes but also that misclassifications will come in the form of common portfolios.

6.1.1 Portfolio Classification Methodology. The secondary model learns to predict the class label based off the probabilities that it belongs to each of the possible labels generated by the first model. Methods for doing this task can be broadly grouped into three categories: portfolio-based, ranking, and classification methods.

Portfolio-based methods use the prediction probabilities from the first model to determine the predicted label on the test data by treating the average prediction probability of each label for each class as a portfolio. Labels are selected based on which portfolio they have the highest similarity with or least distance from. Three metrics were used to measure portfolio similarity: Cosine Similarity (Eq. 5), Manhattan Distance (Eq. 6), and Weighted T-Score (Eq. 7). Cosine similarity finds the most similar portfolio to the sample's prediction probabilities by treating the probabilities and the portfolio as vectors and finding the portfolio whose angle with the probabilities has the highest cosine. Manhattan distance finds the most similar portfolio by finding the one with the smallest sum of the absolute difference of the probabilities for each label. Weighted T-score uses the portfolio's mean and standard deviation for its own label to calculate a T-score for the sample's prediction probability for that class. The similarity measure is the sample's prediction probability for that class divided by the T-score.

$$\arg\max_{A'} \sum_{i=1}^{N} \frac{S_i * P_{A',i}}{\sqrt{\sum_{i=1}^{N} S_i^2}\sqrt{\sum_{i=1}^{N} P_{A',i}^2}} \tag{5}$$

$$\arg\min_{A'} \sum_{i=1}^{N} |S_i - P_{A',i}| \tag{6}$$

$$\arg\max_{A'} \frac{S_{A'} * \sigma_{A',A'}}{|S_{A'} - P_{A',A'}|} \tag{7}$$

Ranking methods compare the order of the probabilities of each label sorted descendingly instead of comparing the exact values. This type of approach helps recognize samples with inflated "nothing" probabilities or shifts in label probabilities that could cause confusion between the portfolios. These ranking methods must also take the prediction probability for the label into account to prevent false positives for labels with low probability that happen to produce the same ranking profile. Two metrics were used to determine the ranking similarity: Thresholded Jaccard Similarity (Eq. 8) and Weighted Jaccard Similarity (Eq. 9). Jaccard similarity

measures the size of the intersection of two sets over the union of those two sets. In this case, the Jaccard similarity is calculated using the top-k, e.g., top-4, predictions of the ranking profile for each class and the sample's ranked label probabilities. Thresholded Jaccard similarity is a piecewise function based on the sample's prediction probability of the class in question. If the probability passes the threshold determined by the mean prediction probability of that class for its own label and the respective standard deviation of that probability, then that label will use the Jaccard similarity of the sample with that class as its metric. Otherwise, that label is given a zero as its metric due to its lack of probability-based similarity to the class. We tested a range of threshold values based on the mean label probability plus or minus the standard deviation and found the mean minus half a standard deviation granted the best balance of precision and recall overall. Weighted Jaccard similarity is Jaccard similarity multiplied by the prediction probability for the label.

$$\arg\max_{A'} \begin{cases} \frac{rank(S) \cup rank(P_{A'})}{rank(S) \cap rank(P_{A'})} & \text{if } S_{A'} > P_{A',A'} - \frac{\sigma_{A',A'}}{2} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$\arg\max_{A'} S_{A'} * \frac{rank(S) \cup rank(P_{A'})}{rank(S) \cap rank(P_{A'})} \quad (9)$$

Classification-based methods simply involve training a classifier using the portfolios as features. The logic behind this is learning what combinations of probabilities from each classifier are seen for each class. As a result, common misclassifications can be identified and corrected. Classifiers used include decision trees, multilayer perceptrons, support vector machines, and random forest.

6.2 Results

The best-performing secondary classifier methods are presented on the user-specific and user-dependent models from Section 5: support vector machine (SVM), multilayer perceptron (MLP), cosine similarity, and Manhattan distance.

6.3 User-Specific Models

With the introduction of an additional classification step after the baseline model, the baseline performance needs to be updated in turn. As shown in Table 6, using the SVM classifier to reclassify samples based off the prediction probabilities of the controlled Bagging classifier significantly improves performance.

6.4 User-Dependent Model

Using the secondary classifier decreased performance on the user-dependent model as seen in Table 7. As it has more training data at its disposal, the user-dependent model generally has higher training performance than the user-specific models. As a result, the prediction probabilities of the training data did not produce patterns that generalized well to the evaluation data. This conclusion is supported by the metrics of cosine similarity and Manhattan distance performing better as these methods are less impacted by the exact values of each probability.

Table 6: User-Specific Models

Participant	SVM	MLP	Cosine	Manhattan
P1	0.699	0.665	0.736	0.74
P2	0.857	0.611	0.563	0.56
P3	0.788	0.796	0.890	0.884
P4	0.646	0.555	0.699	0.67
P5	0.785	0.761	0.679	0.673
P6	0.777	0.739	0.724	0.714
P7	0.794	0.794	0.759	0.749
P8	0.785	0.823	0.685	0.678
P9	0.812	0.813	0.837	0.837
Average	0.771	0.729	0.730	0.723

Table 7: User-Dependent Model

Participant	SVM	MLP	Cosine	Manhattan
P1	0.498	0.498	0.698	0.694
P2	0.700	0.666	0.610	0.590
P3	0.655	0.677	0.683	0.661
P4	0.557	0.557	0.591	0.596
P5	0.650	0.605	0.639	0.634
P6	0.572	0.572	0.693	0.687
P7	0.602	0.646	0.773	0.757
P8	0.547	0.547	0.721	0.778
P9	0.658	0.658	0.842	0.838
Overall	0.604	0.603	0.719	0.710

6.5 User-Adaptive Model

Because the user-specific model is trained only on that one user's data and to learn how the primary classifier tended to misclassify that specific user, the portfolios are calculated just on that user's data. With the secondary classifier, the user-adaptive model reached and barely surpassed the performance of the baseline shown in Table 6 with an average macro F1-score of 0.772.

Table 8: User-Adaptive Model

Participant	SVM	MLP	Cosine	Manhattan
P4	0.670	0.650	0.810	0.795
P1	0.808	0.833	0.608	0.605
P3	0.856	0.844	0.782	0.758
P5	0.654	0.608	0.647	0.647
P7	0.791	0.727	0.724	0.709
P6	0.759	0.827	0.707	0.683
P8	0.744	0.733	0.788	0.775
P2	0.843	0.843	0.722	0.710
P9	0.822	0.828	0.865	0.857
Average	**0.772**	0.766	0.739	0.727

7 DISCUSSION

7.1 Random Forest Results

In addition to Bagging, we tested the performance of the Random Forest algorithm on this classification task. However, the results

were lackluster, barely surpassing an F1-score of 0.6. Due to the extreme data skew, the individual trees most likely favored the majority class and were unable to garner enough votes for the minority class "taking medication" on those samples. These Random Forest results and the realization of why regular Bagging had failed led to the design of the controlled Bagging classifier proposed in this work.

7.2 Interclass and Intraclass Variability

Unlike ambulation activities and other ADLs such as brushing teeth and eating, the action of taking medication not only does not feature characteristic repetitive motion but also inherently features significant interclass and intraclass variability. Individuals can grab a pill bottle with either hand and then remove the cap with the other hand. The pill can then be dumped into the hand not holding the bottle or the person can reach inside the bottle to grab a pill. They can then put down the bottle to take the medication (sometimes with a glass of water) or continue to hold the bottle while taking the medication. The individual may switch which hand is holding the bottle and replace the cap with the opposite hand they opened it with. Furthermore, the individual may not take pills out of the bottle directly at all and instead opt to use other containers such as daily pillboxes or blister packs which have their own unique interactions. This is all to say that each step can be performed in multiple valid ways, and that these choices are not consistent across users or even across different instances of taking medication by the same individual. In our study participants were supplied with a child-proof bottle filled with M&M's to simulate the activity; however, one participant took their own medication during the study which came in a blister pack and a twisting child-proof bottle. We did not instruct participants to take medication in any specific way for this study.

These variations suggest the possibility of breaking the activity up into smaller atomic actions (e.g., picking up a pill bottle with the left hand, picking up a pill bottle with the right hand) and attempting to classify these. This is an approach other studies have taken [12, 27, 38]; however, these studies collect data from a controlled lab environment rather than a real-world setting. Indeed, expecting participants to label their actions at this level of granularity outside of a lab environment would likely result in inaccurate and incomplete labels. A potential solution to this would be to use some sort of clustering technique on taking medication data to find atomic actions. To investigate the feasibility of this we ran a clustering experiment (t-SNE on 10% of the windowed data) to visualize the likelihood that this solution would work. This can be seen in Figure 3. Looking at the visualization it becomes clear that the activity of "taking medication" does not form distinct clusters.

8 LIMITATIONS

While our study was designed to provide us with more realistic data than a study conducted in a laboratory would have, several aspects of our design prevented us from collecting completely natural data. One aspect was having participants take fake medication from a provided pill bottle a number of times over the course of a single data collection session. Future work should focus on training and testing algorithms on data collected from a longitudinal study with individuals taking their own medication as they normally would. Additionally, in our study, participants wore two smartwatches, likely making participants especially cognizant of being studied during data collection and not a practical requirement for real-world deployment of the system. Future work should certainly focus on developing algorithms based on sensor data from a single smartwatch (although the issue of user preference for wristwatch wearing would be present); however, future work should also focus on designing hardware and software that overcomes this limitation. For example, if interfaces leveraging these algorithms were designed for non-wearable devices such as tablets or smartphones, wearable devices could forgo the screen and become comfortable, lightweight bands or bracelets (as is the case for a number of fitness trackers). This could overcome the impracticality inherent to requiring individuals to wear a smartwatch on each wrist.

9 INTELLIGENT MEDICATION ADHERENCE INTERFACES

Automatic detection of medication intake has the ability to change the way medication adherence interfaces are designed in two distinct ways: automating medication management and personalization. Researchers need to go beyond simply recognizing when individuals take medication but also make intelligent decisions on how to best utilize that information. More specifically, algorithms will need to recognize individual-specific medication adherence and non-adherence patterns and subsequently make decisions based on recognition of those patterns to address the individual's actions.

9.1 Automated Medication Management

The current paradigm for medication management largely depends on honest and diligent tracking by individuals. Interfaces for this allow users to perform actions such as create their own schedule, log their medication intake, and set up reminders. However, as soon as the individual stops manually supplying the application with information, the interface becomes ineffectual. Our work; however, could allow interfaces to automate much of this, raising the likelihood that the interface remains effective at ensuring medication adherence. By tracking medication intake habits over time, the system could learn the individual's schedule, intelligently remind them to take their medication only if they have forgotten, and even reliably inform caregivers and/or reassure family members that the individual is taking their medication.

9.2 Personalization

Improvements to medication adherence interfaces could similarly be made in the form of personalization. Systems could learn over time what types of interventions and motivations work best for different individuals; for some simple reminders could be enough, while others might prefer a direct reminder from a family member or caregiver. Gamification approaches could likely become more nuanced. Furthermore, with knowledge of the individual's prescription and calendar, interfaces could provide timely suggestions on when and where to pick up refills.

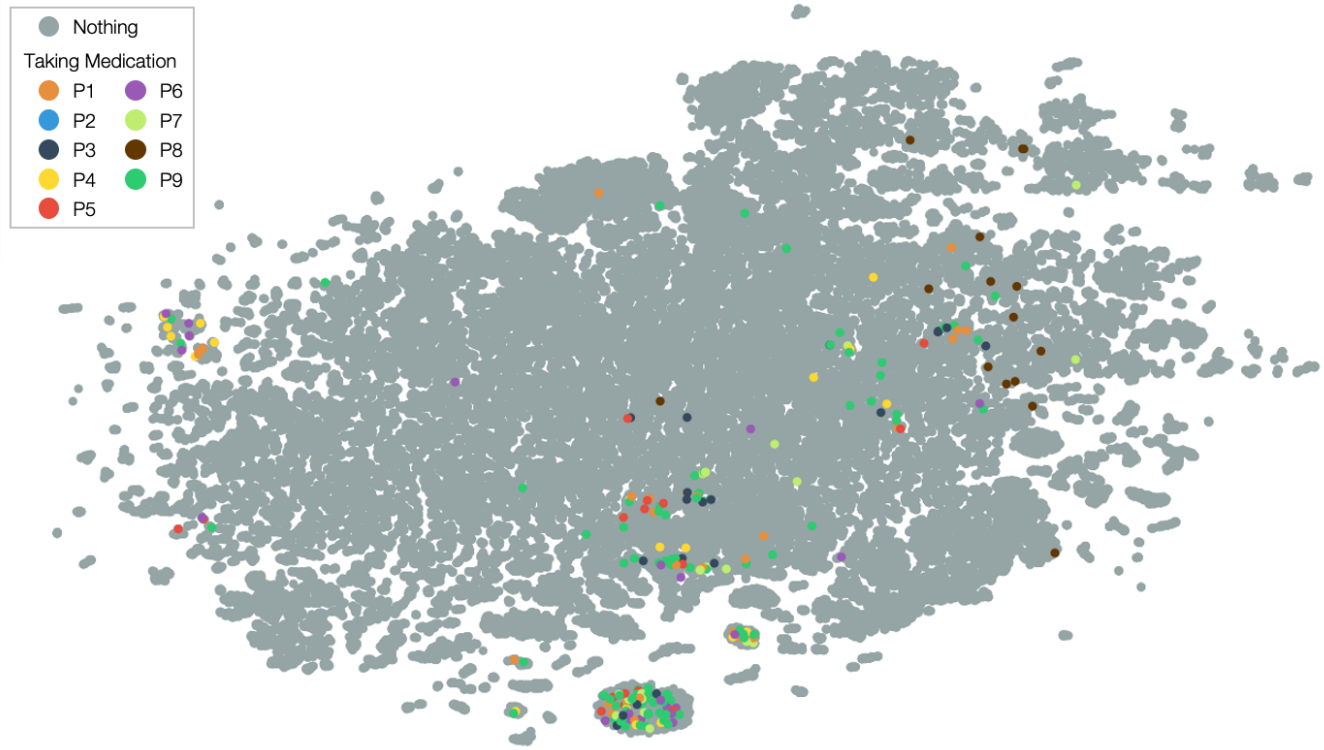

Figure 3: Clustering results from a t-SNE nonlinear dimensionality reduction (perplexity=50, iterations=100000) on 10% of the windowed data. Here "taking medication" is separated by participant, note how neither taking medication over all users and taking medication by user does not cluster.

10 CONCLUSION

Medication non-adherence is a significant and widespread issue that has a number of negative consequences including increased hospital readmission rates, increased healthcare costs, and higher morbidity and mortality rates. Addressing this issue will require the development of flexible proactive health management systems capable of providing timely personalized interventions. In this work we present an activity recognition system capable of recognizing the action of taking medication using accelerometer data collected from a smartwatch. We present portfolio classification as a novel methodology for analyzing the data for instances of taking medication. We find that using a user-adaptive model we are able to recognize when individuals took medication with an F-measure of 0.77. These results show that our system is capable of informing interfaces designed specifically to increase the rate of medication adherence.

ACKNOWLEDGMENTS

We would like to acknowledge the NSF for their support via grant 1952236.

REFERENCES

[1] Zahraa S Abdallah, Mohamed Medhat Gaber, Bala Srinivasan, and Shonali Krishnaswamy. 2018. Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 71.

[2] Folami Alamudun, Hong-Jun Yoon, Kathleen B Hudson, Garnetta Morin-Ducote, Tracy Hammond, and Georgia D Tourassi. 2017. Fractal analysis of visual search activity for mass detection during mammographic screening. *Medical physics* 44, 3 (2017), 832–846.

[3] Murtadha Aldeer, Jorge Ortiz, Richard E Howard, and Richard P Martin. 2019. PatientSense: patient discrimination from in-bottle sensors data. In *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, Houston, Texas, USA, 143–152.

[4] Parviz Asghari, Elnaz Soleimani, and Ehsan Nazerfard. 2020. Online human activity recognition employing hierarchical hidden Markov models. *Journal of Ambient Intelligence and Humanized Computing* 11, 3 (2020), 1141–1152.

[5] Oresti Banos, Miguel Damas, Alberto Guillen, Luis-Javier Herrera, Hector Pomares, Ignacio Rojas, and Claudia Villalonga. 2015. Multi-sensor fusion based on asymmetric decision weighting for robust activity recognition. *Neural Processing Letters* 42, 1 (2015), 5–26.

[6] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*. Springer, Linz and Vienna, Austria, 1–17.

[7] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Granada, Spain, 2546–2554. http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

[8] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] Kam Capoccia, Peggy S Odegard, and Nancy Letassy. 2016. Medication adherence with diabetes medication: a systematic review of the literature. *The Diabetes Educator* 42, 1 (2016), 34–71.

[11] Kyle D Checchi, Krista F Huybrechts, Jerry Avorn, and Aaron S Kesselheim. 2014. Electronic medication packaging devices and medication adherence: a systematic review. *Jama* 312, 12 (2014), 1237–1247.

[12] Chen Chen, Nasser Kehtarnavaz, and Roozbeh Jafari. 2014. A medication adherence monitoring system for pill bottles based on a wearable inertial sensor. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine*

and Biology Society. IEEE, IEEE, New York, NY, USA, 4983–4986.

[13] Liming Chen, Chris D Nugent, and Hui Wang. 2011. A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering* 24, 6 (2011), 961–974.

[14] Josh Cherian, Vijay Rajanna, Daniel Goldberg, and Tracy Hammond. 2017. Did you Remember To Brush? : A Noninvasive Wearable Approach to Recognizing Brushing Teeth for Elderly Care. In *11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, Barcelona, Spain, 48–57. DOI: 10.1145/3154862.3154866, https://dl.acm.org/citation.cfm?id=3154866.

[15] Lindsey E Dayer, Rebecca Shilling, Madalyn Van Valkenburg, Bradley C Martin, Paul O Gubbins, Kristie Hadden, and Seth Heldenbrand. 2017. Assessing the medication adherence app marketplace from the health professional and consumer vantage points. *JMIR mHealth and uHealth* 5, 4 (2017), e45.

[16] Stefan Dernbach, Barnan Das, Narayanan C Krishnan, Brian L Thomas, and Diane J Cook. 2012. Simple and complex activity recognition through smart phones. In *Intelligent Environments (IE), 2012 8th International Conference on*. IEEE, Guanajuato, Mexico, 214–221.

[17] Iram Fatima, Muhammad Fahim, Young-Koo Lee, and Sungyoung Lee. 2013. A Genetic Algorithm-based Classifier Ensemble Optimization for Activity Recognition in Smart Homes. *KSII Transactions on Internet & Information Systems* 7, 11 (2013), 21 pages.

[18] Donya Fozoonmayeh, Hai Vu Le, Ekaterina Wittfoth, Chong Geng, Natalie Ha, Jingjue Wang, Maria Vasilenko, Yewon Ahn, and Diane Myung-kyung Woodbridge. 2020. A Scalable Smartwatch-Based Medication Intake Detection System Using Distributed Machine Learning. *Journal of Medical Systems* 44, 4 (2020), 1–14.

[19] Enrique Garcia-Ceja, Carlos E Galván-Tejada, and Ramon Brena. 2018. Multi-view stacking for activity recognition with sound and accelerometer data. *Information Fusion* 40 (2018), 45–56.

[20] Anil K Gehi, Sadia Ali, Beeya Na, and Mary A Whooley. 2007. Self-reported medication adherence and cardiovascular events in patients with stable coronary heart disease: the heart and soul study. *Archives of Internal Medicine* 167, 16 (2007), 1798–1803.

[21] Katherine A Grosset, Ian Bone, and Donald G Grosset. 2005. Suboptimal medication adherence in Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society* 20, 11 (2005), 1502–1507.

[22] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Andrey Gavrilov, and Sungyoung Lee. 2007. Activity recognition based on semi-supervised learning. In *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2007)*. IEEE, IEEE, New York, NY, USA, 469–475.

[23] Niloofar Hezarjaribi, Ramin Fallahzadeh, and Hassan Ghasemzadeh. 2016. A machine learning approach for medication adherence monitoring using body-worn sensors. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, IEEE, New York, NY, USA, 842–845.

[24] Tâm Huynh and Bernt Schiele. 2006. Towards less supervision in activity recognition from wearable sensors. In *2006 10th IEEE International Symposium on Wearable Computers*. IEEE, IEEE, New York, NY, USA, 3–10.

[25] Aurel O Iuga and Maura J McGuire. 2014. Adherence and health care costs. *Risk management and healthcare policy* 7 (2014), 35.

[26] Haik Kalantarian, Nabil Alshurafa, Ebrahim Nemati, Tuan Le, and Majid Sarrafzadeh. 2015. A smartwatch-based medication adherence system. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, IEEE, New York, NY, USA, 1–6.

[27] H. Kalantarian, N. Alshurafa, and M. Sarrafzadeh. 2016. Detection of Gestures Associated With Medication Adherence Using Smartwatch-Based Inertial Sensors. *IEEE Sensors Journal* 16, 4 (Feb 2016), 1054–1061. https://doi.org/10.1109/JSEN.2015.2497279

[28] Haik Kalantarian, Babak Motamed, Nabil Alshurafa, and Majid Sarrafzadeh. 2016. A wearable sensor system for medication adherence prediction. *Artificial intelligence in medicine* 69 (2016), 43–52.

[29] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[30] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow, Scotland, 1–13.

[31] Oscar D Lara and Miguel A Labrador. 2013. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE* 15, 3 (2013), 1192–1209.

[32] Young-Seol Lee and Sung-Bae Cho. 2014. Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data. *Neurocomputing* 126 (2014), 106–115.

[33] Brent Longstaff, Sasank Reddy, and Deborah Estrin. 2010. Improving activity classification for health applications on mobile devices using active and semi-supervised learning. In *2010 4th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, IEEE, New York, NY, USA, 1–7.

[34] Jinxin Ma, Anaelia Ovalle, and Diane Myung-kyung Woodbridge. 2018. Medhere: A smartwatch-based medication adherence monitoring system using machine learning and distributed computing. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Honolulu, Hawaii, USA, 4945–4948.

[35] Victor Mendiola, Abnob Doss, Will Adams, Jose Ramos, Matthew Bruns, Josh Cherian, Puneet Kohli, Daniel Goldberg, and Tracy Hammond. 2019. Automatic exercise recognition with machine learning. In *International Workshop on Health Intelligence*. Springer, Honolulu, Hawaii, USA, 33–44.

[36] Alexander Mertens, Christopher Brandl, Talya Miron-Shatz, Christopher Schlick, Till Neumann, Andreas Kribben, Sven Meister, Clarissa Jonas Diamantidis, Urs-Vito Albrecht, Peter Horn, et al. 2016. A mobile application improves therapy-adherence rates in elderly patients undergoing rehabilitation: a crossover design study comparing documentation via iPad with paper-based control. *Medicine* 95, 36 (2016), 1–8.

[37] Bobak Jack Mortazavi, Mohammad Pourhomayoun, Gabriel Alsheikh, Nabil Alshurafa, Sunghoon Ivan Lee, and Majid Sarrafzadeh. 2014. Determining the single best axis for exercise repetition recognition and counting on smartwatches. In *2014 11th International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, Zurich, Switzerland, 33–38.

[38] Vasily Moshnyaga, Masaki Koyanagi, Fumiyuki Hirayama, Akihisa Takahama, and Koji Hashimoto. 2016. A medication adherence monitoring system for people with dementia. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Budapest, Hungary, 000194–000199.

[39] Andrea B Neiman, Todd Ruppar, Michael Ho, Larry Garber, Paul J Weidle, Yuling Hong, Mary G George, and Phoebe G Thorpe. 2017. CDC grand rounds: improving medication adherence for chronic disease management—innovations and opportunities. *MMWR. Morbidity and mortality weekly report* 66, 45 (2017), 1248.

[40] Lars Osterberg and Terrence Blaschke. 2005. Adherence to medication. *New England Journal of Medicine* 353, 5 (2005), 487–497.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[42] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. 2005. Activity recognition from accelerometer data. In *AAAI*, Vol. 5. AAAI, Palo Alto, CA, USA, 1541–1546.

[43] Bernardino Romera-Paredes, Min SH Aung, and Nadia Bianchi-Berthouze. 2013. A One-Vs-One Classifier Ensemble With Majority Voting for Activity Recognition.. In *ESANN*. ESANN, Bruges, Belgium, 6 pages.

[44] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications* 59 (2016), 235–244.

[45] Loren J Schleiden, Olufunmilola K Odukoya, and Michelle A Chui. 2015. Older adults' perceptions of e-prescribing: impact on patient care. *Perspectives in health information management* 12, Winter (2015), 15 pages.

[46] Linda Simoni-Wastila, Yu-Jung Wei, Jingjing Qian, Ilene H Zuckerman, Bruce Stuart, Thomas Shaffer, Anand A Dalal, and Lynda Bryant-Comstock. 2012. Association of chronic obstructive pulmonary disease maintenance medication adherence with all-cause hospitalization and spending in a Medicare population. *The American journal of geriatric pharmacotherapy* 10, 3 (2012), 201–210.

[47] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. 2008. Exploring semi-supervised and active learning for activity recognition. In *2008 12th IEEE International Symposium on Wearable Computers*. IEEE, IEEE, New York, NY, USA, 81–88.

[48] Surescripts. 2011. The National Progress Report on e-prescribing and interoperable health care.

[49] Christina Tarantola. 2017. The Top Reminder Apps for Patients. https://www.pharmacytimes.com/contributor/christina-tarantola/2017/12/the-top-medication-reminder-apps-for-patients Last Accessed Sept. 29, 2020.

[50] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Osaka, Japan, 1029–1040.

[51] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. 2016. The MobiAct Dataset: Recognition of Activities of Daily Living using Smartphones.. In *ICT4AgeingWell*. SCITEPRESS, Setúbal, Portugal, 143–151.

[52] Rui Wang, Zdeňka Sitová, Xiaoqing Jia, Xiang He, Tobi Abramson, Paolo Gasti, Kiran S Balagani, and Aydin Farajidavar. 2014. Automatic identification of solid-phase medication intake using wireless wearable accelerometers. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, IEEE, New York, NY, USA, 4168–4171.