Watch-Dog: Detecting Self-Harming Activities From Wrist Worn Accelerometers

Pratool Bharti, Anurag Panwar, Ganesh Gopalakrishna, and Sriram Chellappan

***Abstract*—In a 2012 survey, in the United States alone, there were more than 35 000 reported suicides with approximately 1800 of being psychiatric inpatients. Recent Centers for Disease Control and Prevention (CDC) reports indicate an upward trend in these numbers. In psychiatric facilities, staff perform intermittent or continuous observation of patients manually in order to prevent such tragedies, but studies show that they are insufficient, and also consume staff time and resources. In this paper, we present the Watch-Dog system, to address the problem of detecting self-harming activities when attempted by in-patients in clinical settings. Watch-Dog comprises of three key components—Data sensed by tiny accelerometer sensors worn on wrists of subjects; an efficient algorithm to classify whether a user is active versus dormant (i.e., performing a physical activity versus not performing any activity); and a novel decision selection algorithm based on random forests and continuity indices for fine grained activity classification. With data acquired from 11 subjects performing a series of activities (both self-harming and otherwise), Watch-Dog achieves a classification accuracy of 98%, 94%, and 70% for same-user 10-fold cross-validation, cross-user 10-fold cross-validation, and cross-user leave-one-out evaluation, respectively. We believe that the problem addressed in this paper is practical, important, and timely. We also believe that our proposed system is practically deployable, and related discussions are provided in this paper.**

***Index Terms*—Harmful activity recognition, Random Forest, sensors and public health, smart health.**

I. Introduction

In a survey in 2012, it was reported that psychiatric inpatient suicides account for approximately 5% of the more than 35,000 suicides in the United States [1], with recent reports indicating rising numbers [2]. Psychiatric hospitals most often implement 15-minute manual checks on inpatients (throughout the clock) to prevent suicide attempts. While manual suicide checks are effective at reducing the number of suicides committed by inpatients, studies show that such an approach is ineffective [3], [4]. For instance, the study in [4] shows that out of 15,000 inpatient suicides that were investigated over many years, 20% to 62% of attempts happened when patients were on intermittent observation and 2% to 9% on constant observation [5]. Furthermore, manual checks are known to be prohibitive in consuming nursing resources, non-scalable and have caused other important responsibilities to be overlooked [6]. There is clear need today in hospital settings for alternative or supplementary procedures to combat suicide attempts by inpatients, while also being cost effective.

TABLE I
Activity Set

Normal Activities	Self-harming Activities
Drink with Left Hand (DLH)	Cutting Left Hand (CLH)
Drink with Right Hand (DRH)	Cutting Right Hand (CRH)
Lying Down (LYNG)	Cutting Throat with Left Hand (CTLH)
Running (RUN)	Cutting Throat with Right Hand (CTRH)
Sitting (SIT)	Injection in Left Arm (ILA)
Standing (STND)	Injection in Right Arm (IRA)
Walking (WLK)	Hanging (HNG)
	Smothering (SMTH)

A. Contributions of This Paper

In this paper, we present Watch-Dog, a system with applications in psychiatric facilities to detect self-harming activities (described in Table I) attempts by inpatients. Our system comprises of three components.

1) Wrist-worn Accelerometers: In our system, subjects will have miniaturized accelerometer sensors embedded in accessories worn on both wrists. We chose to have accelerometers in both wrists since the dominant hand can be different for different people when attempting activities. Note that most psychiatric hospitals, and even other healthcare facilities in general provide wristbands with bar-codes for patients to minimize errors in care delivery today. It is straightforward and cheap to embed low cost accelerometers in such bands today as evidenced by recent efforts in the industry and academia [7]. In this paper, for a proof of concept, we utilize the Shimmer[1] wearable sensing device, which is commercially available, energy efficient and widely used. It has an embedded accelerometer, a processing unit and wireless transmission capabilities. The Shimmer device is unobtrusive, and can comfortably be worn on the wrist like a watch. Note though that many commercial wearables like

Manuscript received September 16, 2016; revised March 7, 2017; accepted March 31, 2017. Date of publication April 6, 2017; date of current version May 3, 2018. *(Corresponding author: Pratool Bharti.)*

P. Bharti, A. Panwar, and S. Chellappan are with the Department of Computer Science & Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: pratool@mail.usf.edu; anuragpanwar@mail.usf.edu; sriramc@usf.edu).

G. Gopalakrishna is with the Department of Clinical Psychiatry, University of Missouri, Columbia, MO 65201 USA (e-mail: gopalakrishnag@health.missouri.edu).

Digital Object Identifier 10.1109/JBHI.2017.2692179

[1] http://www.shimmersensing.com/

Microsoft Band and Samsung Gear do provide SDKs to stream real sensory data from their devices for processing, and as such, our technologies in this paper can directly apply when such wearables are worn as well.

2) Algorithm to Determine Active or Dormant State of a Subject: Modern accelerometers are capable of sensing at very high sampling rates. Having a system that continually senses, processes and transmits streaming data from these devices for activity recognition can be energy consuming. In this paper, we improve upon energy and computational overhead by taking advantage of contextual information of inpatients in psychiatric facilities. Specifically, patients in psychiatric facilities are dormant for a significant portion of time (i.e., sleeping, lying down, reading a book etc.), during which time the accelerometer readings are relatively stable. When the patient attempts an activity of interest to this paper, the accelerometer readings will suddenly spike up. By contrasting the spikes in accelerometer readings over a moving window, our proposed technique will effectively determine when a subject is transitioning from a dormant to an active state, and only then will the complex task of fine grained activity classification be attempted. As we show subsequently, this approach results in significant energy savings.

3) Novel Decision Selection Algorithm: Once a subject is determined to be active, the next step is determining the actual activity. To do so, we first employ a Random Forest (RF) based approach to decide on the activity by independently processing the accelerometer readings in each wrist. RF based techniques are fast, accurate and leverage high sampled input data streams with minimal overhead. We then propose an approach that combines the two decisions from either wrist to select one final decision on the activity, based on the notion of continuity indices. In our approach, if there are discrepancies in the activities classified from the accelerometer readings in each wrist, weight is given to that activity whose continuity has been the longest. This intuitive approach improves the accuracy of decision selection with negligible increase in energy expended.

B. Experimental Evaluations

We conducted an experiment with 11 subjects, that was supervised by a clinical psychiatrist. Each subject was instructed to perform a series of 15 activities while a Shimmer device was securely attached to either wrist. Some of the activities were routine (like walking, drinking, etc.) while others were intended to mimic self-harmful behavior (like cutting hand, hanging, etc.). Our proposed techniques achieve an overall classification accuracy of 98%, 94% and 70% for same-user 10-fold cross-validation, cross-user 10-fold cross-validation and cross-user leave-one-out evaluation respectively. The energy expended and latency in decision selection are quite minimal, hence making our system practical.

Note that results of this paper came from an experiment where the subjects were not in a clinic, nor known to have past self-harmful behavior. However, studies do show that suicidal thoughts and tendencies can come without warning, or without advanced planning by those that attempt them [8]. As such, our experimental studies in this paper do have contextual relevance. Nevertheless, we caution against generalizing any conclusions in medical contexts, but rather we demonstrate an innovative application of wearable sensors and activity recognition algorithms in psychiatric facilities, which to the best of our knowledge has not been attempted before. To clarify, more discussions on practical perspectives are presented in the paper.

II. Related Work

A. Activity Recognition Algorithms in HealthCare

In [9] a system is developed to assist physicians to understand patient mobility without direct observation. In this scheme, smartphone accelerometer data collected from patients was used to classify activities like walking, sitting, standing, going upstairs and downstairs. In [10], accelerometer sensors attached on a subject's leg were leveraged to assist patients with Parkinson's disease by detecting episodes where the gait freezes. The system is also designed to send a rhythmic audio signal to stimulate the patient to walk when a freeze happens. In [11], a comprehensive survey is provided on the impact of position of accelerometer sensors on the body for activity classification for healthcare applications.

B. Positioning Our Prior Work w.r.t. This Paper

We have done prior work in detecting self-harming activities from wearable devices in [12]. In [12], we attached smartphones in both wrists for activity recognition via implementing a Dynamic Time Warping (DTW) based algorithm, from accelerometer data. The system in our current paper used Shimmer devices instead, which are much more comfortable to wear like a watch (or an arm-band), and is hence much more practical. In the current paper, our activity recognition framework includes modules for contrasting active/ dormant states of a subject (before attempting fine-grained activity classification) for superior energy efficiency unlike our prior work in [12] that consumed much more energy as a result of performing fine-grained activity classification throughout. We also introduce the notion of continuity indices in current work for superior accuracy via fusing multiple decisions from either wrist unlike [12]. Finally, the evaluation approach in this paper is much more comprehensive by considering three strategies (i.e., same-user 10-fold cross-validation, cross-user 10-fold cross-validation and cross-user leave-one-out evaluation) compared to the work in [12], which was evaluated using only cross-user cross-validation strategy. The current paper also provides significant insights on practical applications of our proposed technologies.

III. The Watch-Dog System: Problem Scope, Hardware Component and Algorithmic Component

In this section, we present in detail our Watch-Dog system for recognizing self harming activities.

Fig. 1. Shimmer devices worn as a watch by a subject.

A. Problem Scope

Our problem is to classify self-harming activities commonly attempted by psychiatric inpatients[2]. Since the system is expected to be operational 24×7 while the patient is in the hospital, minimizing energy and processing latency is vital.

Hanging is the most common form of inpatient suicide [4], requiring only 4 to 5 min to be successful [14]. Other self-harming activities attempted by inpatients also include cutting themselves, hanging and self-injections [15]. After careful discussions with domain experts, a total of eight self-harming activities were identified for detection. To serve as a reference, seven other activities that are not self-harming, but rather ones that patients do as part of daily activities in psychiatric settings were also identified for detection as part of this study. See Table I for full activity list.

B. Hardware Component of our Watch-Dog System

In our system, two Shimmer devices were securely strapped to subjects on either wrist like a watch as shown in Fig. 1. The Shimmer device is widely used in research today for its miniature size and powerful sensing/ computing/ wireless transmission abilities. The central element of the platform is the low-power MSP430F5437A microprocessor with 24 MHz clock rate which controls the operation of the device. The CPU has an integrated 16-channel 12-bit analog-to-digital converter (ADC) which is used to constantly sample and capture tri-axial acceleration signals from an in-built accelerometer in the unit. These accelerometers have a range of ± 16 g (where g is gravitational acceleration) and were sampled at 50 Hz. Note that the frequency of most human activities lie within range of 15 Hz [16]. As such, a sensor sampling rate of 50 Hz is ideal for our problem, since according to the Nyquist rule for loss-less reconstruction of a signal, it needs to be sampled at a rate that is at-least twice its highest frequency [17].

To achieve synchronization from units in both wrists, data was recorded using Shimmer Sync software, that synchronizes time stamp data from both accelerometers. Devices were calibrated using standard calibration techniques as described in [18]. Subsequently, the accelerometer readings from both Shimmer device were streamed via an in-built bluetooth radio module within the unit to a computer for post-processing.

C. Algorithmic Components of Watch-Dog System

In this section, we elaborate on the algorithmic components of the Watch-Dog system. In our implementation for this paper, we point out that our algorithms execute on a computer where data from both Shimmer devices are streamed (via bluetooth) for activity classification.[3]

The algorithmic framework of Watch-Dog is shown in Fig. 2. Accelerometer data from each Shimmer device is independently pre-processed to first remove noise. Then, in order to determine whether a subject or active or dormant, it is fed into the STA/ LTA module (discussed in Section III-C2). Once the subject is determined to be active, our system performs feature extraction from accelerometer data coming from each wrist, and attempts to classify the corresponding activity independently using a Random Forest based algorithm. Then, the decision identified from each wrist is integrated using the notion of continuity indices to determine the final activity. Each step is explained in detail below.

1) Data Pre-Processing: The first step of our algorithmic framework is pre-processing the raw accelerometer data from the Shimmer device in each wrist. Depending on the orientation of Shimmer device, gravity can influence the readings on one or more of the components. To avoid this issue, Shimmer API provides methods to sample linear acceleration directly and hence eliminating the influence of gravity. Once the linear acceleration data is extracted, we further pre-process it by applying a median filter to smooth the data and remove any unexpected spikes [19]. We experimented through all odd numbers of samples from length 3 to 31 and finally set the length to 21 to get good smoothing on accelerometer signal. Further we feed the data to a low pass filter using a 15Hz cut-off 4th order Butterworth filter to limit the bandwidth of the signal to the frequencies common in human motion, hence removing high frequency noise.

Once noise is removed and signal is smoothened, the next step is to determine an appropriate sliding window size for the signals to attempt run-time classification. A window size of $WS = 200$ accelerometer samples from either wrist (approximately 4 seconds) with 50% overlap was used to create a new database, W, that was used as the training/testing data for activity classification. In prior related work in [20] it is found that 2-5 seconds window works best for human activity recognition using accelerometer data. Hence, we conducted our experiment with window length from 2 to 5 seconds having 0.5 second intervals and found that window length of 4 seconds is working best for our problem. Subsequently, the segmented window W is forwarded to next steps for activity classification. An snapshot of accelerometer readings from a Shimmer device in the left hand for various activities is shown in Fig. 3 for visualization.

2) STA/ LTA Triggering Algorithm: Once the pre-processed readings from both accelerometers are ready, the data is fed into the STA/ LTA processing module to determine if the

[2]Two highest-risk times for suicide for psychiatric inpatients are in the week after admission and very shortly after discharge [13].

[3]It is very easy to also implement our framework on a smartphone as well.

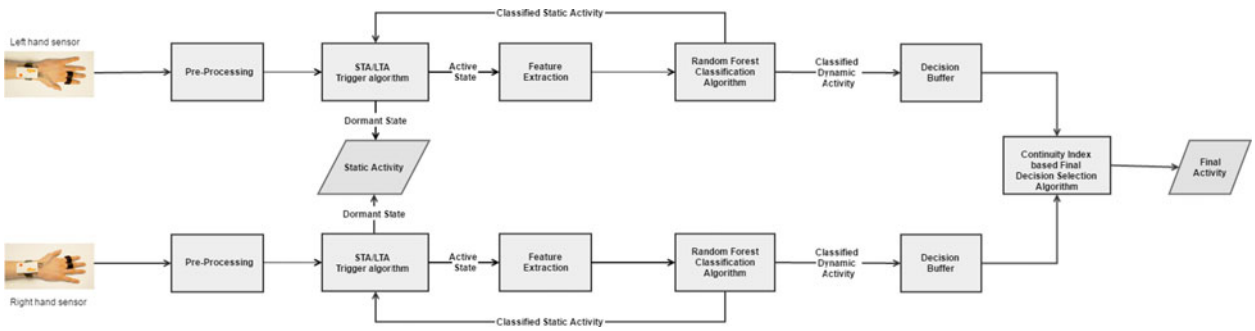

Fig. 2. Algorithmic framework of Watch-Dog system.

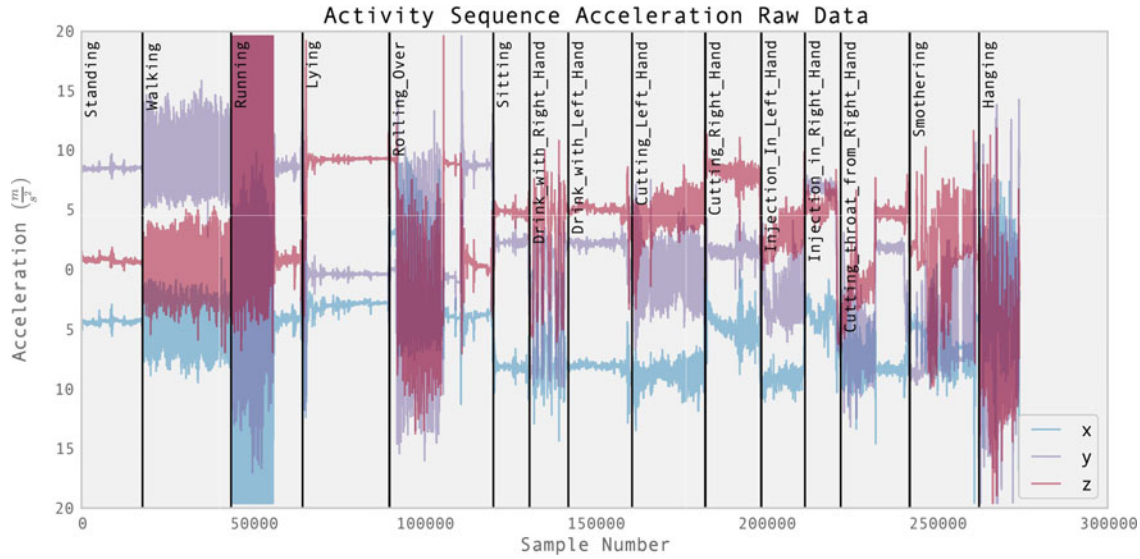

Fig. 3. Recorded accelerometer readings throughout an activity sequence from the Shimmer device secured to participants left hand. The blue series denotes the x component, red shows y, and purple shows z.

subject is active or dormant. The Short-Time-Average/ Long-Time-Average (STA/ LTA) algorithm is an algorithm used in seismology to detect sudden spikes in vibration for earthquakes detection [21]. The algorithm is simple, and energy efficient, and is applicable to our problem scope in psychiatric settings, where patients are dormant for a significant portion of time wherein the accelerometer readings are stable, and when the patient is active, the accelerometer readings suddenly spike up. In this manner, as long as a subject is dormant, no fine-grained activity classification will be attempted by our system. Rather, only when a subject is determined to be active our framework will attempt the more complex task of activity classification, hence saving energy.

In Figs. 3 and 4, we illustrate our rationale pictorially for a subject. In Fig. 3, we can visually see that the accelerometer readings for activities like standing, sitting and lying down are stable, compared to the more dynamic activities. We further quantify this in Fig. 4, where we see that variance in accelerometer readings (for just the x axis) computed for dormant activities like standing, lying and sitting is very low, while the variance for other activities is higher. How to leverage this insight to detect

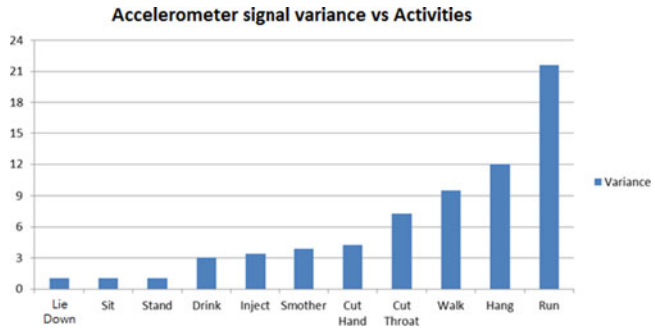

Fig. 4. Comparison of accelerometer variance of dormant vs active states.

transitions from dormant to active states within the context of our problem scope is our challenge.

When we process the variance to detect activity transitions, there are a few parameters to consider. The first is L_a, which integrates acceleration variance in three axes in a moving Long Term Window. The Long Term Window (T_{lta}) is a moving window over a long time frame that captures the long-term stability

Algorithm 1: STA/ LTA Algorithm.

Data: STA Window length ($T_{sta} = 0.5$ second), LTA Window length ($T_{lta} = 15$ seconds), Tri-axial accelrometer data (a_{xt}, a_{yt}, a_{zt}), Trigger threshold (TR_{th}) = 2.5

Result: Active vs Dormant States

while *True* **do**

$S_a = 1 + Var[a_{xt}]_{t=1}^{T_{sta}} + Var[a_{yt}]_{t=1}^{T_{sta}} + Var[a_{zt}]_{t=1}^{T_{sta}}$

$L_a = 1 + Var[a_{xt}]_{t=1}^{T_{lta}} + Var[a_{yt}]_{t=1}^{T_{lta}} + Var[a_{zt}]_{t=1}^{T_{lta}}$

if $S_a/L_a > TR_{th}$ **then**

Activities classified as $Active$ state;

else

Activities classified as $Dormant$ state;

Continue;

end if

end while

in accelerometer readings when the subject is dormant. The second parameter is S_a, which integrates acceleration variance in three axes in a moving Short Term Window. The Short Term Window (T_{sta}) is a moving window over a short time frame that captures the short term spikes in accelerometer readings when the subject is transitioning from a dormant to an active state. Finally, the last parameter is a ratio denoted as $TR_{th} = \frac{S_a}{L_a}$, which is compared against an predetermined application-based threshold value to determine dormant or active state of a user.

In our system, values of these parameters are constant. Therefore, they need to be set carefully. We experimented with many different set of values and finally set the T_{sta}, T_{lta} and TR_{th} as 0.5 second, 15 seconds and 2.5 which gave us perfect discrimination between dormant and active states for activities of interest to this paper. To arrive at these values, we tested $TR_{th} = \frac{S_a}{L_a}$ for every pair of activities. The ratio for transitions from a dormant activity to another dormant activity varied from 0.9 to 1.2 and the ratio for a dormant activity to a active state activity varied from 3 to 21. Therefore, to detect transitions from dormant to active states, setting TR_{th} as 2.5 is an ideal choice for our problem scope. The working structure for STA/ LTA is shown in Algorithm 1. As we can see, the accelerometer readings are sensed continuously, and processed via the STA/ LTA module. When the ratio of the short term variances in acceleration and the long term variances in acceleration crosses the threshold of 2.5, the subject is classified as active now, and the process of fine-grained activity classification begins, and is discussed next.

3) Our Algorithm for Activity Classification: In this section, we present our algorithm for classifying activities once a subject is determined to be active from the STA/ LTA module. Core requirements of our system are accuracy, fast response, and ability to handle high sampled data streams. In our system, our algorithm is based on the notion of Random Forests (RF) [22]. Basically, RF is an ensemble supervised learning technique, wherein multiple lightweight decision trees are constructed, and the algorithm searches multiple trees for probabilistic classification. It is fast and accurate (since multiple light-weight trees are constructed), and handles streaming sensor data very well. It is also energy efficient, as we demonstrate later in performance evaluations. High data storage may be an issue for RF if data is big and number of trees in forest are high. Since our algorithm runs on moderately configured system with modest forest size and limited data, storage is not an issue.

Feature Extraction & Selection: Feature extraction and feature selection from input data are critical for any supervised learning algorithm. Too few features may not be representative, and too many features incur processing overhead and sometimes can even decrease accuracy by introducing noise [23]. As such, it is critical that we identify a limited set of features from accelerometer data that provide good discriminatory power among various activities of interest, while also keeping processing delay and energy low.

To start with, we extracted 200 features from the accelerometer readings that were intuitive and used in past studies within our problem scope[4]. From this vast set, we applied Wrapper-based feature selection algorithm [24] to select the most relevant features out of one representative feature subset from all features. Since wrappers are more fine tuned towards a classifier, they generally achieve high classification accuracy. We applied wrapper based feature selection method on our training dataset using Random Forest classifier to evaluate subsets by their predictive accuracy (on test data) by statistical re-sampling or cross-validation. As a result, we selected 12 best features from this pool including both time and frequency domain. All 12 features are listed in Table II. These features serve as an input vector x into Random Forest algorithm for activity selection.

Random Forest Algorithm Design: Random Forest (RF) is a supervised learning algorithm. It is a voting based ensemble of L decision trees (DT). Each DT works as a independent classifier and predicts one activity from processing that particular tree. The final activity selected from the algorithm is the one selected by a majority of the trees.

A DT is represented as $\{T_i(x, \theta_i)\}$, where x is an input feature vector extracted from raw accelerometer data and θ_i is a random vector that dictates the structure of ith tree. The random vector θ_i is generated independent of the preceding θ_1 .....θ_{i-1} vectors, but with the same distribution. In the random subspace method, θ_i consists of a K integers (K$\ll$ M) randomly drawn from a uniform distribution in the interval [1, M], where M is the number of available features. Given a dataset set that contains N feature vectors, each consisting of M features, the RF algorithm builds the trained model using following steps:

1) Draw N samples at random with replacement from the dataset, to generate the training set of the tree.
2) Select any K features randomly from the set of available features, where K$\ll$ M.
3) Among the values for each of the K features drawn, choose the best binary split according to the Gini impurity index [25], which measures impurity degree in dataset. Gini index value lies between 0 and 1. It is maximum when all classes in dataset have equal probability and minimum when any one class has maximum prob-

[4]Due to space limitations, we do not present them in the paper.

TABLE II
FINAL FEATURES SELECTED FROM POOL OF FEATURES

$Norm = \sum_{i=1}^{N} \sqrt{(a_{xi})^2 + (a_{yi})^2 + (a_{zi})^2}$
$Standard\ Deviation = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \mu)^2}$
$Max = \underset{i \in \{1,2..N\}}{\operatorname{argmax}}\ (a_i)$
$Min = \underset{i \in \{1,2..N\}}{\operatorname{argmin}}\ (a_i)$
$Entropy = -\sum_{i=1}^{N} p_i (\log p_i)$
$Autoregressive\ Parameters = \sum_{i=1}^{N} AR_{param}\, a_{i-1} + \varepsilon_i$
$Correlation = \frac{\sum_{i=1}^{N}(a_{xi}-\bar{a}_x)(a_{yi}-\bar{a}_y)}{\sqrt{\sum_{i=1}^{N}(a_{xi}-\bar{a}_x)^2 \sum_{i=1}^{N}(a_{yi}-\bar{a}_y)^2}}$
$Max\ reduced\ Mean = (\underset{i \in \{1,2..N\}}{\operatorname{argmax}}\ a_i) - \bar{a}$
$No\ of\ peaks = \quad Count\ of\ local\ maxima$
$Spectral\ energy = \sum_{f=0}^{fs/2} \lvert a[f] \rvert^2$
$Maximum\ Frequency = \underset{i \in \{1,2..N\}}{\operatorname{argmax}}\ FFT(a_x, a_y, a_z)$
$Mean\ absolute\ Deviation = \frac{\sum_{i=1}^{N} \lvert a_i - \mu \rvert}{N}$

ability. Finally select those features which has the least impurity.

4) Grow the tree to its maximum size according to the stopping criterion chosen and let the tree unpruned.

Once the forest has been ensembled, an unseen data sample is labeled with one of the activity classes having the maximum conditional probability summed up over all decision trees: i.e., it is labeled with the activity which has maximum probability combined by probability from each ensemble trees. In the RF approach, given a feature sample x to be classified, the conditional probabilities for each activity are computed by taking the average of the conditional probabilities given by the trees constructing the ensemble. These conditional probabilities are computed as follows. Given a decision tree T, and an input feature sample x to be classified, let us denote by $v(x)$ the leaf node where x falls when it is classified by T. The probability $P(a|x,T)$ that the sample x belongs to the activity a, where $(a \in A_1, A_2, ..., A_{15})$, is estimated by the following equation:

$$P(a|x,T) = \frac{n_a}{n} \tag{1}$$

where n_a is the number of training samples falling into $v(x)$ after learning and n is the total number of training samples assigned to $v(x)$ by the training procedure. Given a forest consisting of L trees and an unknown feature sample x to be classified, the probability estimate $P(a|x)$ that x belongs to the activity a is computed as follows:

$$P(a|x) = \frac{1}{L}\sum_{i=1}^{L} P(a|x,T_i) \tag{2}$$

where $P(a|x,T_i)$ is the conditional probability provided by the ith tree and is computed according to (1). As a consequence, for the sample x to be classified, the RF algorithm gives as output the vector:

$$\mathbf{p} = \{P(A_1|x), P(A_2|x) \ldots\ldots P(A_{15}|x)\} \tag{3}$$

The activity with the highest probability in the set is chosen as the final classified activity for the entire ensemble forest [26]. The workflow of the Random Forest algorithm with pre-processing, training and testing phase is formally shown in Algorithm 2.

Recall that in our system, accelerometer readings from each Shimmer device on either wrist will be be processed as above to determine an activity. Once this is completed, we have two activities independently identified one from each wrist-worn device. Integrating activities from both hands to decide on the final activity is the last step and is discussed next.

4) Final Decision Selection Algorithm: Algorithm IV-B presents our final decision selection algorithm that integrates the individual decisions from each wrist-worn accelerometer. As we show later in Section IV, in a majority of cases, the activity classified from both hands is the same because of the effectiveness of our random forest approach. However, there is a chance that this may not happen, and the activities identified by each hand may be different. This usually happens when there are some unexpected dynamics in one or more hands during performing of an activity that confuses our algorithm. To address this issue in a simple, intuitive and energy efficient manner, we introduce the notion of continuity indices for final activity selection.

In this technique, a small buffer table is used for both Shimmer devices separately. The buffer table holds activity predictions (generated from our Random Forest algorithm) from recent past segmented windows from each device. We define a new term called continuity index as the number of times the current activity predicted appeared consecutively in buffer table. Any activity of interest to this paper is continuous in time. For instance, subjects are extremely unlikely to sleep in one moment, and start drinking the very next moment, and move on to another activity immediately. This property is true for all activities of interest to our problem. As such, if decisions on final activity from either wrist are different, we give preference to that activity which has been continuously detected the longest from either wrist. This method effectively helps eliminate the impact of outliers affecting our final decision, and is also simple and energy efficient to implement.

IV. RESULTS FROM EXPERIMENTAL EVALUATIONS

In this section, we present results of experimental evaluation of our system. We first present the data collection process, then the metrics, and finally the results of our evalu-

Algorithm 2: *(Continued)*

Training data from right and left sensors $= TD_r, TD_l$;
Testing data from right and left sensors $= D_r, D_l$;
Features extracted from right and left sensors $= F_r, F_l$;
Classified Activity from right and left sensors$= A_r, A_l$;
Prob. that feature F belongs to Activity $A = P(A|F)$;
Segmented window size $= W$;
No. of trees in Random Forest $= L$;
Step 1 Pre-Processing:
1) Median filters are applied to remove accidental spikes from D_r, D_l.
2) Low-pass filters are applied to remove high frequency signals from D_r, D_l.
3) Features F_r, F_l are extracted from processed data D_{rp}, D_{lp} obtained from (1) and (2).
Step 2 Training:
Input: Training data set F_r, F_l
Output: Random Forest model to classify normal vs harmful activities
1) Select a bootstrap sample of size N from the training data.
2) Grow a decision tree T by selecting K features at random from the set of M features. Choose the best feature among the K. Split the node into two daughter nodes and let the tree grow to its maximum size.
Step 3 Prediction:
Input: D_r, D_l
Output: A_r, A_l
while *True* **do**
if Window size $> W$ **then**
if STA/ LTA algorithm triggers $Active$ state **then**
$F_r, F_l =$ Extracting feature set from D_{rp}, D_{lp};
for *each* T *in* $Forest$ **do**
$$P(A|F) = \frac{1}{L}\sum_{i=1}^{L} P(A|F,T);$$
end for
$$A_r = \underset{i\in\{1,2..15\}}{\operatorname{argmax}} \Big(P(A_i|F_r)\Big)$$
$$A_l = \underset{i\in\{1,2..15\}}{\operatorname{argmax}} \Big(P(A_i|F_l)\Big)$$
if A_r and A_l are available and valid **then**
$A_{fs} =$ **Final Activity Selection**(A_l, A_r)
end if
else
$A_{fs} =$ Static or Safe activity
end if
end if
end while
Data: A_l, A_r; Detected activity from right and left sensors
Result: Finally activity selected; A_{fs}
Step 4 Final Activity Selection:
Initialization;
Left hand buffer table $= LHB_{i,i-1,..i-h}$;

Algorithm 2: *(Continued)*

Right hand buffer table $= RHB_{i,i-1,..i-h}$;
Finally activity selected $= F_{as}$;
Current selection $= i$;
Size of buffer table $= h$;
while decision is available from both sensors **do**
if $i < h$ **then**
LHB· Add[$A_l(i)$];
RHB· Add[$A_r(i)$];
else
if $A_l(i)$ is equal to $A_r(i)$ **then**
$F_{as} = A_l(i)$;
else
for $k = h; k >= 0; k - -$ **do**
if $A_l(i)$ is equal to $A_l(i-k)$ **and** $A_r(i)$ is not equal to $A_r(i-k)$ **then**
$F_{as} = A_l(i)$;
break;
else if $A_l(i)$ is not equal to $A_l(i-k)$ **and** $A_r(i)$ is equal to $A_r(i-k)$ **then**
$F_{as} = A_r(i)$;
break;
else
$F_{as} =$ Activity from Dominant hand;
break;
end if
end for
end if
end if
end while

ation. In our experimental studies, a total of 11 adult subjects were recruited for the study, and the data collected was split for training and testing as is standard in machine learning techniques.

A. Data Collection

In our experiment, two Shimmer devices were securely strapped to subjects on either wrist like a watch as shown in Fig. 1. All subjects that participated in the experiment attested that the devices were un-obtrusive. In our experiment, a clinical psychiatrist supervised all experiments to subjects. The supervisor informed each subject to maintain an activity for approximately three minutes before initiating a new activity. The activities of interest to this paper were presented earlier in Table I. Sequences of activities did not follow any specific order as long as each one was performed by the subject. However, the subjects performed all activities one after the other in a continuous sequence. In Fig. 5, the continuous sequence of activities for two subjects is shown for reference. The Shimmer device was programmed such that the accelerometer readings from each unit was exported in real-time via bluetooth to a computer, where the data was immediately tagged with the corresponding activity using a tagging application developed in $C\#$.

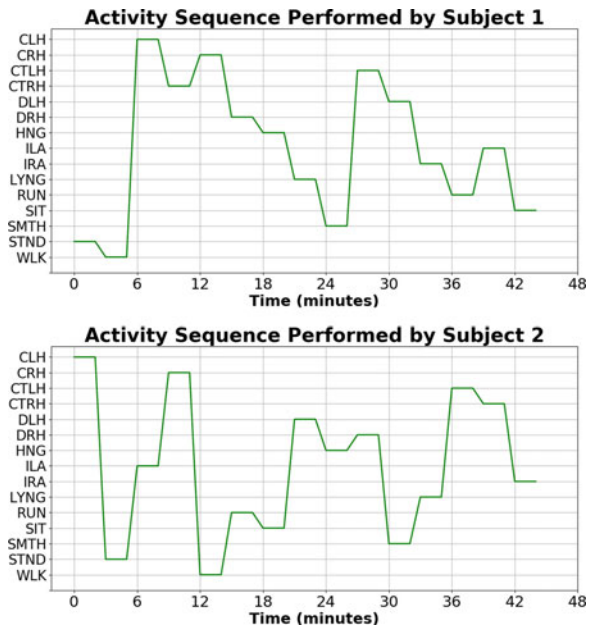

Fig. 5. Sequence of activities performed by Subject 1 and Subject 2 while data collection.

For each subject, an average of 9000 accelerometer samples (for 3 minutes with 50Hz frequency) in each axis (x, y and z) were collected for each activity. A total 89 windows were extracted from each activity and each window consists of $4\,\mathrm{s}$ seconds of data with 50% overlapping which is subsequently used for training and testing.

Note that the training was conducted only for Random Forest Algorithm because STA-LTA module does not need training to function. Also, the training and testing procedures happened offline. The testing dataset was randomized to remove any bias towards evaluating STA-LTA and Random Forest algorithms effectively. The algorithms were executed offline for the results reported below.

B. Metrics

The results of Watch-Dog are presented in terms of *accuracy* and *Confusion Matrix*. Accuracy metric is a function of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The *accuracy* of a classifier is the overall classification performance defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4}$$

The *Confusion Matrix* (CM) is a specific table layout that allows visualization of the performance of a supervised learning algorithm. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa) [27].

The confusion matrix reflects about how confused a prediction model is. For example if an activity is predicted correctly only 40% of the time, then this matrix will show how the algorithm confused its prediction with the other (wrongly classified) activities the remaining 60% of the time.

C. Results

Overview of Evaluation Methods: In this paper, we evaluate the performance of our system using three well established methods that are standard for our problem scope. These testing methods are same user 10-fold cross-validation, cross user 10-fold cross-validation and cross user leave-one-out cross-validation.

10-fold cross-validation divides the dataset into 10 subsets, and evaluates them 10 times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together to form a training set. Then, the average error across all 10 trials is computed for final result. Within this method, there are two approaches to evaluate. In the *Same user 10-fold cross-validation* method the data evaluated belongs to only one subject. In *Cross user 10-fold cross-validation* method, the data is aggregated from all subjects and then 10-fold cross-validation is applied. In *Cross user leave-one-out* method out of n subjects, $n-1$ are chosen for training dataset and one is left for testing. The process repeats for every subject then average is computed for final result.

While discussing results few things are very important to point out. First, the distribution of each activity in dataset is kept uniform. This removes inherent biases, and yields results that are fair. Also, among the three strategies evaluated, some may show better results that others. Usually evaluations on same users show better result compared to any cross user technique. This is intuitive, since there are subtle variations among people even when they do the same activity that are sometimes hard to detect when training and testing are done on different subjects. However, as we show, our algorithms still achieve high performance both within and across users for a number of activities, hence demonstrating the effectiveness of our system. However, with more training and testing across more subjects, we clearly expect improved outcomes. Also, due to space limitations, we do not present evaluations of the STA/LTA module to detect the active or dormant state of a subject, since the algorithm performed with 100% accuracy every time in our experiments.

Results and Interpretations: At the outset, we point out that Watch-Dog obtained 98%, 94% and 70% overall accuracy for same user cross-validation, cross user cross-validation and cross user leave-one-out cross-validation testing methods respectively. The Standard Deviation is also shown in the figures where appropriate.

In Figs. 6 and 7, we show the performance of our system for *same user 10-fold cross validation method*. In Fig. 6 (Top), we present our activity classification results in the form of accuracy. The X-axis refers to the activity classified (identified in Table I), and the Y-axis refers to the performance in accuracy percentage. In Fig. 6, for each activity three accuracy metrics is shown in the corresponding legends. The top, middle and last figures show the accuracy from same user 10-fold, cross user 10-fold and cross user leave-one-out cross-validation evaluation strategy. The corresponding Confusion Matrix for *same user 10-fold cross validation method* is presented in Fig. 7. Similarly in Figs. 6 (middle) and 7 (middle), we show the performance

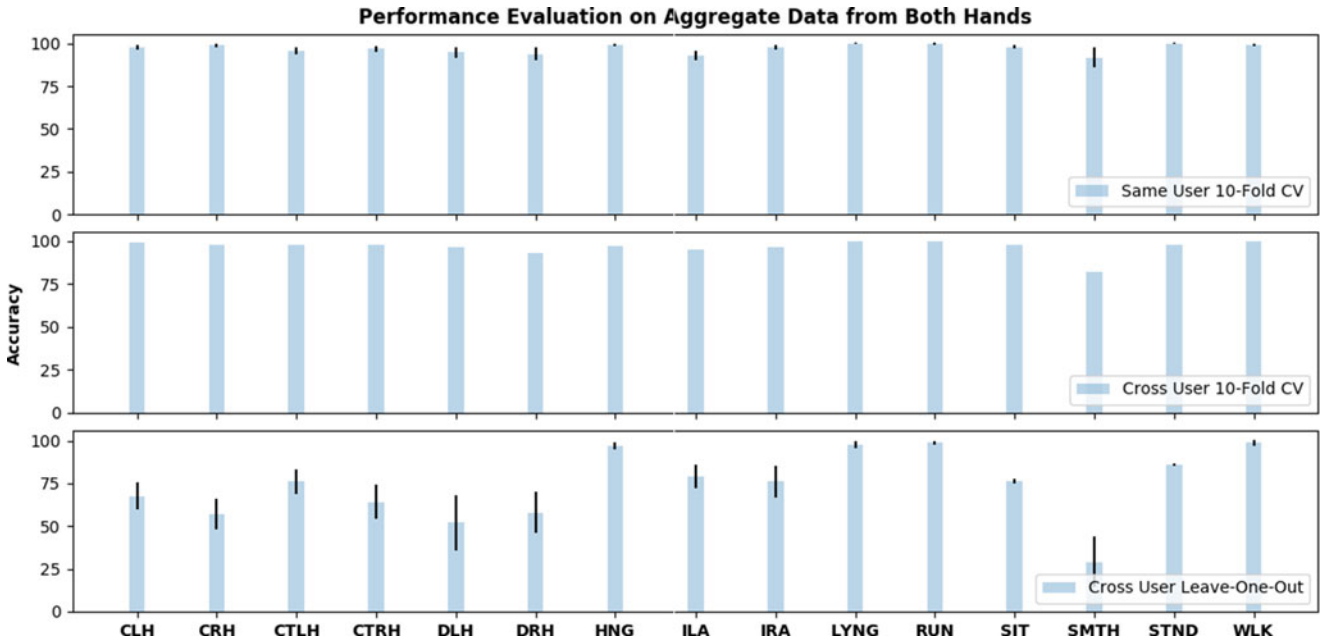

Fig. 6. Accuracy matrix evaluated on aggregated data from both hands.

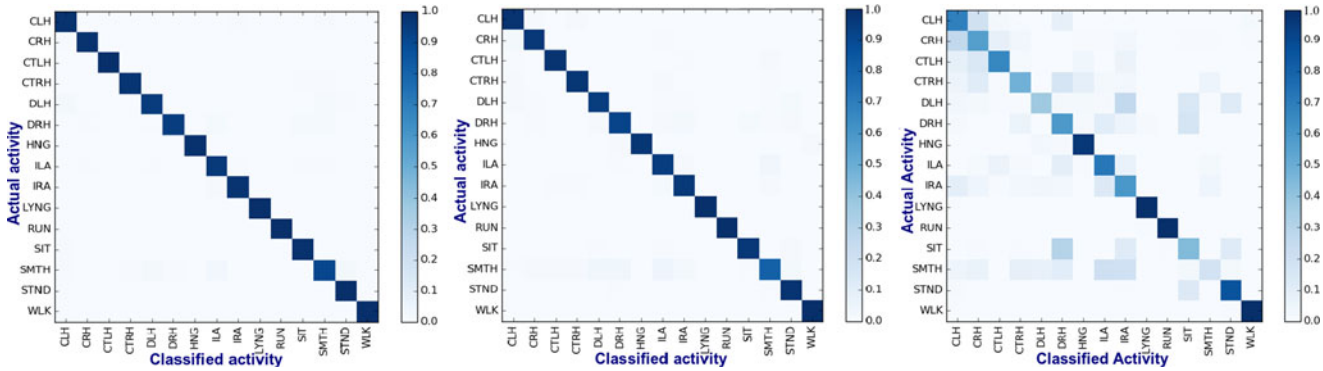

Fig. 7. Confusion matrix for evaluation by Same user 10-fold cross-validation method (left), Cross user 10-fold cross-validation method (middle), Cross user Leave-one-out cross-validation method (right).

(in terms of Accuracy) and the Confusion Matrix for the *Cross user 10-fold cross-validation* method.

As we can see, the performance of our algorithms in terms of accuracy across all activities is very high. This demonstrates the validity of the Random Forest approach for classifying self-harming activities. Furthermore, we can see that the accuracy is still very high when the activity is attempted to be classified independently from the right or left wrist. As such, the improvement in these two evaluation strategies is minor with the Continuity Index approach that integrates decisions from both wrists. This is further validity of our Random Forest approach for activity classification for our problem scope of detecting self-harming activities.

In Fig. 6 (Bottom), we demonstrate the performance of our system for cross-user leave one out strategy, which is the stricter benchmark, since not only the testing data sets are completely unseen to training data sets unlike the above evaluation strategies but also testing subject was not allowed to give samples for training data. We can see in this case, that independent decisions from either wrist are not so accurate like in the previous evaluation strategies, and the need for our Continuity Index approach is more prominent here. For some activities like *Cutting* and *Drinking*, the improvement is as much as 50%, which is quite significant.

The overall performance in terms of accuracy is about 70% for this evaluation method, which is lower than ideal. While adding more data from more subjects will help improve the system, from the confusion matrix in Fig. 7 (Right), we can see that some our system confuses some self-harming activities with others - for instance *cutting* with *drinking*, *injecting with right hand* with *injecting with left hand*. It also confuses some non self-harming activities with other non self-harming activities - like for example *standing* with *sitting*. This is because, when evaluated across users, the subtle differences

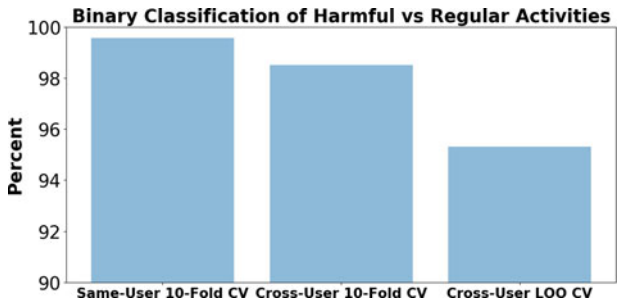

Fig. 8. Accuracy for binary classification of Self-harming vs non Self-harming activities.

in performing self-harming activities confuse the system more than expected. But when we see other activities like running, walking, lying down they have more similar pattern from one subject to another. That is the reason for the overall relatively lower performance in this stricter evaluation strategy.

A Note on Binary Classification: We agree that for a system like ours, and in the relatively sensitive (psychiatric) settings where they are intended for, false negatives and false positives are important parameters. High false positives in detecting self-harming activities will burden staff, and false negatives are more dangerous for patients. To address this fact, we present results of our framework, where the problem is not fine-grained activity classification, but rather a binary one of detecting if an activity performed in self-harming or not. This problem, which we discussed with domain experts is very important in psychiatric facilities, since accurately detecting that a self-harming activity is being attempted by an inpatient may itself enough to trigger an urgent response from the healthcare staff to save the patient in most cases, and the need for actually determining the fine-grained activity (while also important) may be secondary. We present results for the binary classification problem in Fig. 8, where the accuracy of detection is excellent - in fact it is 95% even for the leave-one-out stricter evaluation strategy, which is very encouraging. Building upon this result, and further understanding of domain specific challenges for superior activity recognition in psyciatric facilities is part of on-going work.

Discussions on Energy Efficiency and Latency: Energy evaluation is very crucial for Watch-Dog system due to its nature of running continuously 24×7. We conducted our evaluation on a system with Intel Core i7-5600 processor having clock frequency of 2.60 GHz. We ran STA-LTA and RF independently for half an hour each to test CPU stress by keeping other settings intact. For executing the STA-LTA module, the CPU usage was always under 2% with average below 1% whereas the peak for executing Random Forest with the Continuity Index algorithm during activity classification was under 10% with average below 5%. The average delay incurred while executing our algorithms from start to finish for one window of data was around 500 ms–1.5 s, which demonstrates the speed of our proposed system. Note that on an average, the STA/ LTA module takes about 300 ms to make a decision, while the RF algorithm take about a second to predict the class on an input window. These numbers are quite reasonable, hence enabling the practicality of our contributions. While the execution of our algorithms in this paper was done offline, we are currently designing a framework to enable the system operate in real-time. Also, implementing all algorithms in the form of an executable smartphone app, and evaluating its overhead is part of on-going work.

V. Conclusion

In this paper, we have presented Watch-Dog, a system to detect self-harming activities with applications in psychiatric hospital facilities. Watch-Dog comprises of wrist worn accelerometers, algorithms to detect active or dormant state of a subject, and fine grained classification algorithms to detect self-harming activities. We demonstrated the performance of our system from several metrics and also with multiple evaluation strategies. To the best of our knowledge, ours is the first work that addresses a problem related to activity recognition with core applications to aid inpatients in psychiatric hospitals. Considering this, we also highlighted important practical perspectives of our Watch-Dog system.

Acknowledgment

This work was supported in part by the US National Science Foundations under Grants CNS 1205695 and IIS 1559588. Any opinions, thoughts and findings are those of the authors and do not reflect views of the funding agency.

References

[1] Y. M. Jabbarpour and G. Jayaram, "Suicide risk: Navigating the failure modes," *Focus*, vol. 9, pp. 186–193, 2011.
[2] Online, "Injury prevention & control: Division of violence prevention," Centers Disease Control Prevention, 2015. [Online]. Available: http://www.cdc.gov/violenceprevention/pdf/suicide-datasheet-a.pdf
[3] K. A. Busch, J. Fawcett, and D. G. Jacobs, "Clinical correlates of inpatient suicide," *J. Clin. Psychiatry*, vol. 64, no. 1, pp. 14–19, 2003.
[4] L. Bowers, T. Banda, and H. Nijman, "Suicide inside: A systematic review of inpatient suicides," *J. Nervous Mental Dis.*, vol. 198, no. 5, pp. 315–328, 2010.
[5] J. Knowles, "Inpatient suicide: Identifying vulnerability in the hospi tal setting," 2012. [Online]. Available: http://www.psychiatrictimes.com/suicide/inpatient-suicide-identifying-v ulnerability-hospital-setting
[6] G. Jayaram, H. Sporney, and P. Perticone, "The utility and effectiveness of 15-minute checks in inpatient settings," *Psychiatry (Edgmont)*, vol. 7, no. 8, pp. 46–49, 2010.
[7] Z. D. Gonzalez, "The new wave of wristbands," www.wearable-technologies.com, 2014. [Online]. Available: https://www.wearable-technologies.com/2014/01/the-new-wave-of-wristbands/
[8] P. J. Skerrett, *Suicide Often not Preceded by War*. Cambridge, MA, USA: Harvard Health Publications, 2012.
[9] S. L. Lau, I. Konig, K. David, B. Parandian, C. Carius-Dussel, and M. Schultz, "Supporting patient monitoring using activity recognition with a smartphone," in *Proc. 2010 7th Int. Symp. Wireless Commun. Syst.*, 2010, pp. 810–814.
[10] M. Bächlin *et al.*, "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, Mar. 2010.
[11] M. J. Mathie, A. C. Coster, N. H. Lovell, and B. G. Celler, "Accelerometry: Providing an integrated, practical method for long-term, ambulatory monitoring of human movement," *Physiol. Meas.*, vol. 25, no. 2, pp. R1–R20, 2004.
[12] L. Malott, P. Bharti, N. Hilbert, G. Gopalakrishna, and S. Chellappan, "Detecting self-harming activities with wearable devices," in *Proc. 2015 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2015, pp. 597–602.

[13] P. Qin and M. Nordentoft, "Suicide risk in relation to psychiatric hospitalization: Evidence based on longitudinal registers," *Arch. Gen. Psychiatry*, vol. 62, no. 4, pp. 427–432, 2005.
[14] P. Burgess, J. Pirkis, J. Morton, and E. Croke, "Lessons from a comprehensive clinical audit of users of psychiatric services who committed suicide," *Psychiatric Serv.*, vol. 51, no. 12, pp. 1555–1560, 2000.
[15] E. D. Ballard, D. Henderson, L. M. Lee, J. M. Bostwick, and D. L. Rosenstein, "Suicide in the medical setting," *Joint Commission J. Qual. Patient Safety/Joint Commission Resour.*, vol. 34, no. 8, pp. 474–481, 2008.
[16] R. San-Segundo, J. M. Montero, R. Barra-Chicote, F. Fernndez, and J. M. Pardo, "Feature extraction from Smartphone inertial signals for human activity segmentation," *Signal Process.*, vol. 120, pp. 359–372, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016516841500331X
[17] H. Landau, "Sampling, data transmission, and the Nyquist rate," *Proc. IEEE*, vol. 55, no. 10, pp. 1701–1706, Oct. 1967.
[18] F. Ferraris, U. Grimaldi, and M. Parvis, "Procedure for effortless infield calibration of three-axial rate gyro and accelerometers," *Sensors Mater.*, vol. 7, no. 5, pp. 311–330, 1995. [Online]. Available: http://porto.polito.it/1404539/
[19] C. Randell, C. Djiallis, and H. Muller, "Personal position measurement using dead reckoning," in *Proc. 2003 7th IEEE Int. Symp. Wearable Comput.*, Oct. 2003, pp. 166–173.
[20] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.
[21] A. Trnkoczy, "Understanding & setting STA/LTA trigger algorithm parameters for the k2," 1998.
[22] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[23] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proc. 9th Nat. Conf. Artif. Intell.*, vol. 91, 1991, pp. 547–552.
[24] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.
[25] J. L. Gastwirth, "The estimation of the Lorenz curve and Gini index," *Rev. Econ. Statist.*, vol. 54, no. 3, pp. 306–316, 1972.
[26] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer-Verlag, 2005, pp. 165–192.
[27] Wikipedia, "Confusion matrix—Wikipedia, the free encyclopedia," 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=70448 7108. Accessed on: Feb. 17, 2016.