mORAL: An *m*Health model for inferring Oral Hygiene Behaviors in-the-wild using wrist-worn inertial sensors

SAYMA AKTHER, University of Memphis
NAZIR SALEHEEN, University of Memphis
SHAHIN ALAN SAMIEI, University of Memphis
VIVEK SHETTY, University of California, Los Angeles
EMRE ERTIN, The Ohio State University
SANTOSH KUMAR, University of Memphis

We address the open problem of reliably detecting oral health behaviors passively from wrist-worn inertial sensors. We present our model named *mORAL* (pronounced *em oral*) for detecting brushing and flossing behaviors, without the use of instrumented toothbrushes so that the model is applicable to brushing with still prevalent manual toothbrushes. We show that for detecting rare daily events such as toothbrushing, adopting a model that is based on identifying candidate windows based on events, rather than fixed-length timeblocks, leads to significantly higher performance. Trained and tested on 2,797 hours of sensor data collected over 192 days on 25 participants (using video annotations for ground truth labels), our brushing model achieves 100% median recall with a false positive rate of one event in every nine days of sensor wearing. The average error in estimating the start/end times of the detected event is 4.1% of the interval of the actual toothbrushing event.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*;

Additional Key Words and Phrases: mHealth, brushing detection, flossing detection, hand-to-mouth gestures

ACM Reference Format:
Sayma Akther, Nazir Saleheen, Shahin Alan Samiei, Vivek Shetty, Emre Ertin, and Santosh Kumar. 2019. mORAL: An *m*Health model for inferring Oral Hygiene Behaviors in-the-wild using wrist-worn inertial sensors . *ACM Comput. Entertain.* 9, 4, Article 39 (March 2019), 24 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Healthcare spending is projected to siphon off nearly 20 percent of the United States economy by 2026 [1]. The unsustainable nature of this spending has prompted a growing shift from reactive and expensive healthcare focusing on treating illnesses to proactive, preventive approaches for tackling the underlying health behaviors that increase the risk of diseases. In this work, we focus on dental diseases (e.g., caries and periodontal diseases), a common chronic disease with considerable repercussions [Benjamin 2010]. In the United States, half of

[1]https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsProjected.html

Authors' addresses: Sayma Akther, sakther@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Nazir Saleheen, nsleheen@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Shahin Alan Samiei, ssamiei@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Vivek Shetty, vshetty@ucla.edu, University of California, Los Angeles, Los Angeles, California, 90095; Emre Ertin, ertin.1@osu.edu, The Ohio State University, Columbus, Ohio, 43210; Santosh Kumar, skumar4@memphis.edu, University of Memphis, Memphis, Tennessee, 38152.

1544-3574/2019/3-ART39 $15.00
https://doi.org/0000001.0000001

adults suffer from periodontal diseases and over fifty-three million people live with untreated tooth decay in their permanent teeth; significant subset of this population endure the debilitating consequences of advanced periodontal diseases [CDC et al. 2010]. Beyond the pain and suffering, oral health problems affect the ability to eat and swallow, speak and socialize, and provoke local and systemic infections — health corollaries that greatly diminish an individual's general health and sense of well-being and exact substantial personal and societal costs. Poor oral hygiene is also linked to heart diseases, stroke, diabetes, pneumonia, other respiratory diseases, pre-term births, and low birth-weight babies [Yap 2017].

Ironically, dental diseases are largely avertible and closely linked to the inadequate performance of simple oral hygiene behaviors such as regular toothbrushing and flossing. The American Dental Association (ADA) recommends that everyone should brush their teeth at least twice daily and floss their teeth at least once. Yet, research has found that a significant percentage of the population does not follow these recommendations; 33% of men brush only once a day, and 59% of women regularly skip brushing at bedtime [Chadwick et al. 2011]. This disparity between health recommendations and practice led clinical researchers to argue for development and dissemination of mHealth approaches to assist users in optimizing salient oral health behaviors and thus, reduce the public health burden of dental diseases [Shetty et al. 2018].

For optimal oral health, brushing should last at least two minutes, cover each tooth surface adequately using optimal pressure, and be augmented by flossing. The industry has introduced feature-rich smart toothbrushes to assist users in tracking their brushing behaviors. Smart toothbrushes can track brushing duration and provide user feedback regarding pressure applied to the teeth during brushing [Grossman et al. 1996]. Researchers have explored adding newer sensors for improving the detection of brushing surfaces [Huang and Lin 2016], instrumented miniature cameras into the toothbrush head for detecting plaque [Yoshitani et al. 2016], and developed implantable assistive brushing devices [CHIIZ [n. d.]; Li et al. 2013] to assist children and individuals with disabilities.

While these advancements help users of smart toothbrushes, the vast majority of the population (over 80% in the United States[2]), continue to use manual toothbrushes and will not benefit from these advancements. Inferring the timing, duration and quality of brushing behaviors using manual toothbrushes will bring these benefits to the masses and set the stage for personalized individual and population oral health management, disease risk stratification, hybrid health insurance programs [Shetty et al. 2018], personalized feedback, and individual engagement by providing rewards or gamification [Chang et al. 2008; Nakajima et al. 2007].

Finally, the automated inference of brushing and flossing behaviors can be used in a wide variety of research studies to discover predictors of dental disease outcomes and inform treatments as well as public health policies.

Accomplishing this vision requires a robust computational model for reliably detecting toothbrushing and flossing behaviors in the natural field environment. We use inertial sensors in wrist-worn devices for this purpose that are already being used today to detect sleep, activity, eating, smoking, and several other daily behaviors.

1.1 Challenges

Inferring oral hygiene behaviors (OHBs) primarily from inertial sensors on the wrist presents several technical challenges.

Variability in sensor mounting: Placement of the sensor in a wristband or the position of the wristband on the wrist itself can vary between devices, and even for a given device worn by a specific individual, between wearing episodes (e.g., either palm-facing or back-palm-facing). Each configuration (see Figure 3) produces significantly different signals. Because the signal from the axis parallel to the hand is essential to determine whether the hand is facing in the upward direction or not, the model needs to determine the inertial sensor's configuration relative to the wrist.

[2]https://www.statista.com/statistics/278116/us-households-usage-of-manual-toothbrushes/

Reliable detection of rare daily behaviors: Brushing and flossing are salient and relatively transient events that can take only four minutes out of approximately 960 awake minutes per day. In general, events that take less than 1% of the assessment time have stringent requirements of recall and even more stringent requirements for controlling false positives. Even a 1% false positive rate can produce 4 false positive events per day.

Access to fine-grained accurate labels for model training/evaluation: Obtaining fine-grained labels of each brushing event is necessary to train and validate the OHB models. Specifically, one needs the precise markings (at second-level granularity) of the start and end of each event, and also the start and end times of any intervening pauses. Such carefully labeled data is usually difficult to collect from the free-living natural environment.

Precise estimation of duration and start/end times: In order to be clinically useful, the detected event must match closely the duration and start/end times of the actual event.

1.2 Contributions

Our work presents the *mORAL* (pronounced *em oral*) model for reliably detecting brushing (with manual uninstrumented toothbrushes) and flossing behaviors from inertial sensors worn on the wrists. It makes several contributions. First, we propose a solution to the sensor mounting problem. Second, we propose an explainable method for reliable detection of rare daily behaviors (e.g., applicable to brushing, flossing, rinsing, eating, drinking, smoking, etc.) that achieves greater computational efficiency and detection accuracy by reducing the amount of data to be assessed by an order of magnitude. We show that the event-based approach to identifying candidate segments of data on which to apply a machine learning-based model significantly outperforms the fixed window-based approach. Third, we propose metrics for reporting the error in estimating the start/end times of detected events.

1.3 Organization

Section 2 presents related works. Section 3 describes the data collected and annotation of video for labeling the events for model training and testing. Section 4 presents all steps in model development, Section 5 describes their implementation, and Section 6 presents their evaluation. Section 7 presents limitations and future works; Section 8 concludes the paper.

2 RELATED WORKS

Wrist-worn inertial sensors have been used for detecting a wide variety of Activities for Daily Living (ADL) such as walking, sleeping, eating, combing hair, dressing, climbing stairs, sitting, standing, and cooking [Bruno et al. 2014, 2012, 2013; Ellis et al. 2014; Pavey et al. 2017; Zhang et al. 2012]. Some of these [Bruno et al. 2012; Hsieh et al. 2016] have demonstrated the feasibility of detecting brushing from hand gestures in the context of detecting a vast amount of ADL activities. Due to their focus on covering a large number of ADL's, they trained their models on data collected mostly in scripted settings. As their focus was on demonstrating feasibility, the false positive rate for each class, in particular for toothbrushing, was more than 15%. Such models are not usable in the field for passive detection as they would result in a large number of (i.e., 72) false positives each day (see Table 1).

For a more reliable model that can work on continuous data collected passively in the natural field setting, researchers have been developing models specific to the target behavior. For example, [Thomaz et al. 2015] focused on eating and [Parate et al. 2014; Saleheen et al. 2015]) on smoking. Behaviors such as toothbrushing, smoking, and eating are transitory events that last for only a few minutes. But, continuous data collection through an entire day of wearing produces 16 hours of data (corresponding to an awake portion of the day). Hence, the accuracy requirements for both recall and false positive are stringent. For example, a model for detecting brushing with a 5% false positive rate will produce 24 false positive events per day (see Table 1).

Table 1. The table shows the number of false positive events produced per day by a model for specific false positive rates. It is assumed that sensors are worn for 16 hours per day, and the toothbrushing event lasts an average of 2 minutes.

False positive rate	**15%**	**10%**	**5%**	**1%**	**0.1%**	**0.01%**
False positives per day	72	48	24	4	$\frac{1}{2}$	$\frac{1}{20}$

The major obstacle in using existing models for detecting brushing or flossing is that different behaviors require the development of behavior-specific models. For example, the model for detecting physical activity like walking [Mannini et al. 2013; Pavey et al. 2017] is not directly applicable to detecting eating. Similarly, the models for eating [Thomaz et al. 2015] or smoking [Saleheen et al. 2015] are not directly applicable to detecting brushing or flossing. Hence, reliable detection of brushing passively from wrist-worn sensors is still an open problem.

As described earlier, most of the efforts to detect OHBs have focused on instrumented or smart toothbrushes. Some researchers have utilized commercial smart toothbrushes [Chang et al. 2008; Kim et al. 2009; Lee et al. 2012] while others instrumented manual toothbrushes by either adding inertial sensors [Huang and Lin 2016] or attaching a mini smartphone [Bozgeyikli et al. [n. d.]]. A focus of these works has been to recognize the tooth surface being brushed [Chang et al. 2008; Huang and Lin 2016; Lee et al. 2012; Marcon et al. 2016]. Some research teams have utilized innovative acoustic-based approaches where audio data, collected by a smartphone application or a microphone are inserted into the brush head to evaluate toothbrushing performance [Hachisu and Kajimoto 2012, 2015; Korpela et al. 2016]. [Korpela et al. 2015] analyzed the audio signals of brushing strokes via machine learning techniques to recognize surfaces. In contrast, [Ouyang et al. 2017] introduced a tooth brushing monitoring system that combines a microphone placed on the neck with an earphone placed in the ear.

In general, the investigations using smart or instrumented toothbrushes do not provide a reliable solution for automatically detecting toothbrushing or flossing events from wrist-worn inertial sensors because they rely on users activating a button (as is the case in smart toothbrushes) to indicate the start and end of brushing activity. But, they represent important complementary works that can be applied to the case of manual toothbrushes once a model to detect brushing and flossing events is developed that also identifies the start/end times of these events. The sensor data corresponding to the brushing and flossing event can be used subsequently for deeper analysis of brushing surface detection, pressure monitoring, and for developing engagement apps that improve oral hygiene.

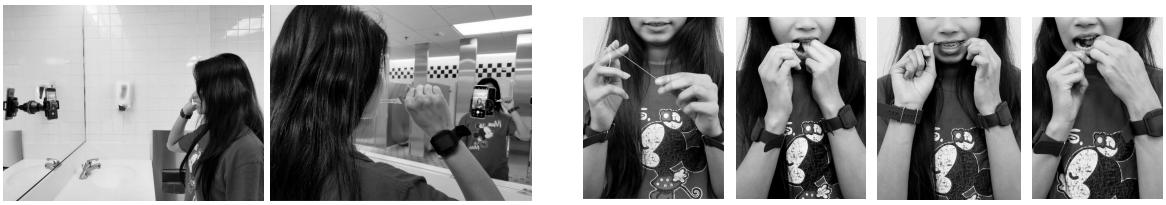
3 DATA COLLECTION

To develop and test our mORAL model, we conducted a field study to collect labeled sensor data in the natural environment of participants. The study was approved by the Institutional Review Board (IRB) and all participants provided written informed consent.

3.1 Devices and Sensor Measurements

The study included three devices carried by each participant:

(1) Participants wore a MotionSense ("wristband") on each wrist (left and right) that included 3-axis accelerometers sampled at 16 Hz and 3-axis gyroscopes sampled at 32 Hz. Data was transmitted from these sensors in real-time using Bluetooth Low Energy (BLE) radio connection to a smartphone.
(2) Each participant carried a study provided Android smartphone. The smartphone was used to communicate with, receive, and timestamp data from the sensor suites. The phone also collected data via its internal sensors, including 3-axis acceleration, 3-axis gyroscope, GPS traces (geo-location data), battery state, and user interaction data. Participants also used the study smartphone's front-facing camera to record videos



(a) Brushing video data collection setup

(b) Dental flossing with string

Fig. 1. Participants used the study smartphone to record videos of themselves each time they performed an OHB (e.g., toothbrushing or flossing) while wearing the wristbands. The video was annotated to obtain labels for model training and testing.

of themselves while performing their oral health routines, including brushing of teeth, flossing, and/or using an oral rinse.

(3) A commercially available, Bluetooth-enabled OralB 6000 toothbrush ("SmartBrush") was provided as compensation for participation in the study. Participants were asked to use their personal (manual) toothbrush once daily, the SmartBrush once daily, and to floss at least once daily, while wearing all the sensors.

3.2 Software for Data Collection

The open-source mCerebrum software [Hossain et al. 2017] was installed on the study smartphone to collect and store the sensor data streams as well as well to initiate video recording by participants. Because time synchronization between the video capture and wristband data was required to correct for any clock drift, mCerebrum was modified to maintain accurate time synchronization of 200 milliseconds.

3.3 Study Protocol

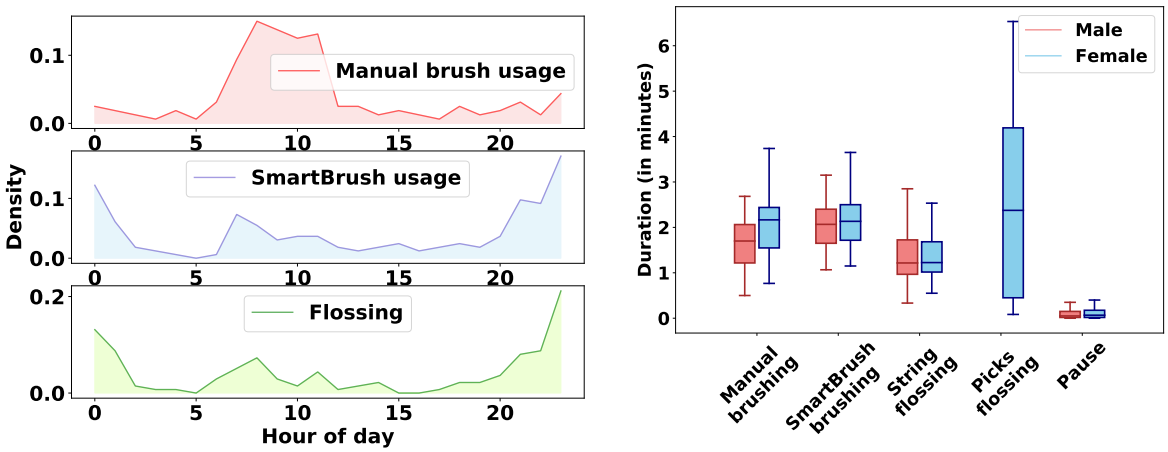
At the time of recruitment, potential study participants were asked about their willingness to brush at least twice daily and floss at least once daily. Once enrolled, participants wore the wristbands during their waking hours for seven (7) days. Participants were asked to use the study smartphone daily to record videos of themselves when performing their typical oral health routines (e.g., toothbrushing or flossing) in a comfortable setting (e.g., at home), while wearing the wristband (Figure 1).

3.4 Video data collection

The videos provided the ground truth labels for toothbrushing and flossing events. The phone stored these data on an encrypted microSD card and periodically uploaded the sensor data (except video and GPS) via a dedicated HTTPS connection to a secure server [Hnat et al. 2017]. Video and GPS data were recovered from the phone at the end of the study period.

3.5 Annotation of OHBs from video data

To create labeled data for model development and evaluation, we annotated the video data to record the timing of each OHB. In total, 362 videos were collected, each with an average duration of 3.12 minutes. We annotated the brushing type (normal or smart-brush) and flossing type (string or picks), and their start/end times from the videos. A key challenge was to mark the pause segments between and within each brushing and flossing event so that for model training and testing, only data corresponding to active brushing or flossing were used.



(a) (b)

Fig. 2. (a) Time of day distribution for brushing and flossing events. Participants usually brushed their teeth with manual toothbrush in the morning and with SmartBrush at night; (b) Duration of oral hygiene events obtained from video annotations.

For brushing, special attention was paid to the interval between the lowering and raising of the brush-holding hand from the mouth. This interval was marked as a pause. In each video, in addition to marking the start and end of each brushing, flossing, and all pause events contained within the brushing and flossing events, we also marked the orientation configuration of both wrists, the brushing wrist (left or right), use of manual or SmartBrush, flossing with string or picks, flossing wrist (left or right), and video pause times.

All the video ground truth data were labeled by two coders. Any discrepancy in their coding was resolved via joint viewing of the segment in doubt, and a consensus was reached regarding the labeling of the event in consideration. A summary of coding definitions appears in Table 3 (see Appendix A).

3.6 Dataset description

A total of 25 participants (12 males, 13 females; mean age 28.5 ± 7.6 years) completed the study. Over this period, 2,797 hours of sensor data were collected over 192 days. A total of 160 brushing events with the manual brush and 164 brushing events with the SmartBrush were observed during this period. A total of 137 flossing events were recorded. Of these, one-third of the flossing events were associated with normal brushing and two-third were associated with SmartBrush. In terms of modality, 81% of flossing were performed using string, and the remaining using picks.

3.7 Descriptive video observations

We use the video annotations of OHBs to make several descriptive observations. First, we analyzed the time of day when OHBs occurred (see Figure 2a). We observe that OHBs occurred across most times of the day, with higher concentrations in the morning and at night. We observe a higher occurrence of manual brushing in the morning and SmartBrush use at night. We also observe that flossing frequency is higher at nighttime.

Second, we analyzed the duration of different OHB events (see Figure 2b). We observe that even though participants were video recording themselves, the duration of brushing with a manual toothbrush was marginally lower than that with SmartBrush. The brushing duration with SmartBrush has a median duration of 2.1 minutes. The SmartBrush vibrates to indicate 2 minutes-worth of brushing, and thus encourages compliance with the

ADA-recommended brushing duration. We also observe that females had a marginally longer brushing duration when brushing with a manual toothbrush. Finally, we observe that flossing occurred for a median duration of 1.22 minutes, and pause durations were very short.

4 MODEL DEVELOPMENT

Before describing our modeling approaches to reliably detect brushing (with a manual toothbrush) and flossing, from wrist-worn inertial sensors, we introduce some notations and definitions.

4.1 Notations and Definitions

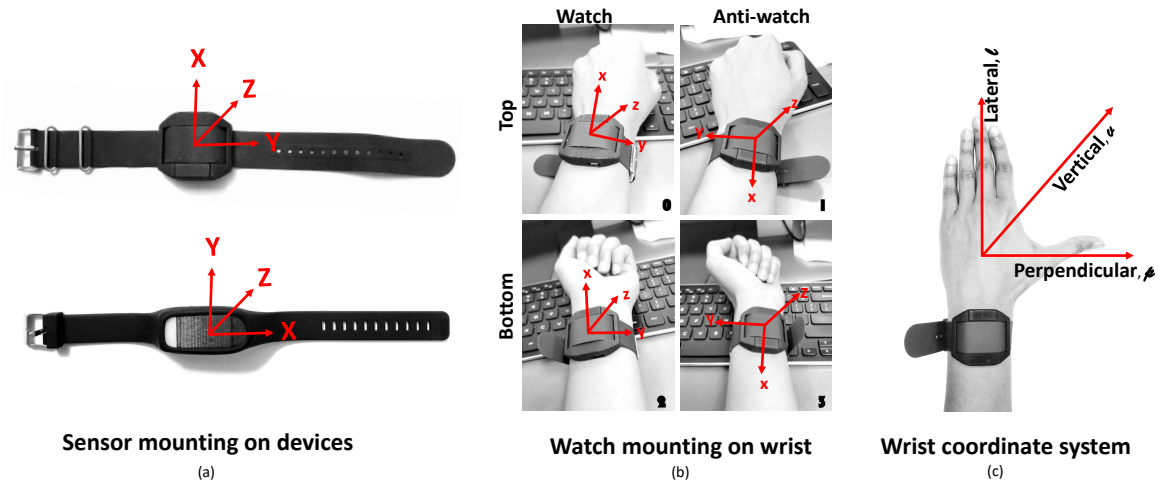


Fig. 3. (a) Lateral (l), perpendicular (p), and vertical (v) axes of wrist coordinate system; (b) Variation in sensor mounting on the wrist-worn devices (c) Four sensor positions on the wrist, referred to as Configuration c (for $c \in \{0, 1, 2, 3\}$)

Sensor data: Let $a_s(t) = (a_x, a_y, a_z)$ be accelerometer data and $\omega_s(t) = (\omega_x, \omega_y, \omega_z)$ be gyroscope data at time t.

Wrist coordinate system: Different wrist-worn devices have different placements of the triaxial inertial sensors. For example, y-axis in one sensor may correspond with the x-axis in another sensor (see Figure 3a). Similarly, the same device can be worn in different ways (see Figure 3b). Therefore, models for inferring wrist orientation and hand gestures, need to be independent of the sensor placement.

To accomplish this, we define a new coordinate system, called *the wrist coordinate system*. The three axes in the wrist coordinate system are defined as lateral (l), perpendicular (p), and vertical (v) axes. Here, lateral axis is aligned with the arm, perpendicular axis is aligned with the thumb, and vertical axis is the gravity axis when the palm is parallel to the earth's surface (see Figure 3c). We denote $a(t) = (a_l, a_p, a_v)$ to be the corresponding accelerometer data in the wrist coordinate system and $\omega(t) = (\omega_l, \omega_p, \omega_v)$ to be the corresponding gyroscope data in the wrist coordinate system.

Sensor orientation configurations: The configuration of a wrist-worn sensor is defined as the current orientation of the sensor relative to the wrist. In other words, it specifies the mapping between sensor's axes coordinate and the wrist coordinate system. Usually, the z-axis is perpendicular to the surface of the sensor. Hence, we assume that the vertical axis is always aligned with the z-axis. For example, in Figure 3b, mapping for the top left placement is $(l, p, v) = (x, y, z)$. We regard this as the *base* configuration, since it occurs most frequently.

4.2 Key Observations

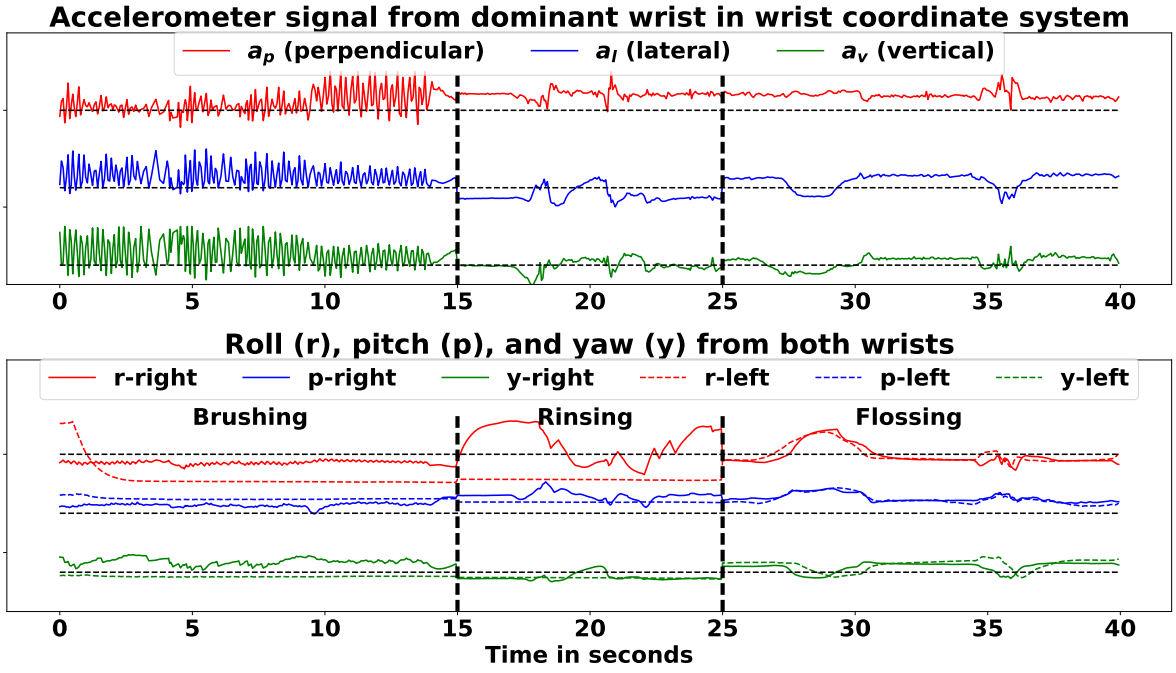


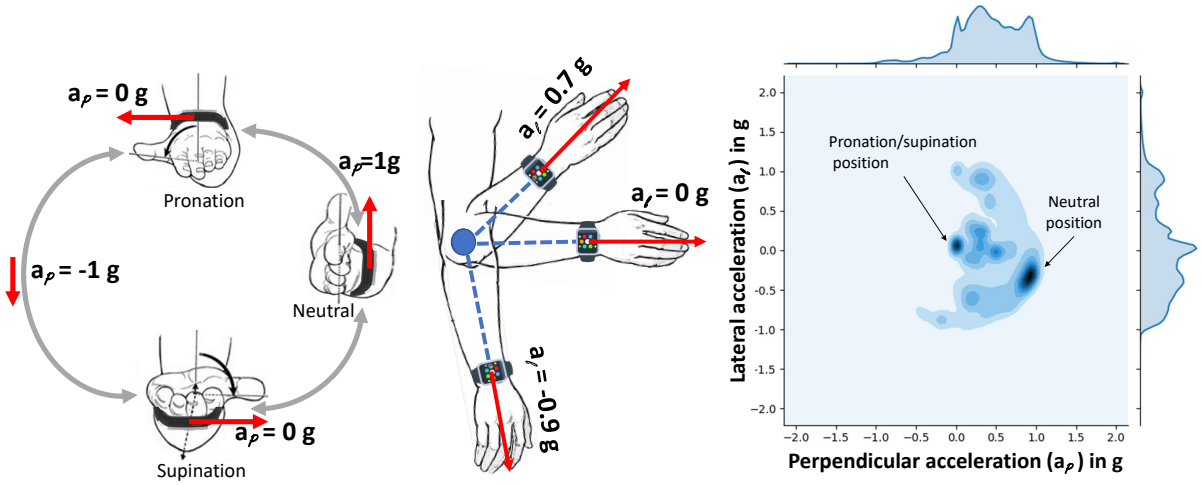
Fig. 4. Wristband signals during brushing, rinsing and flossing with string. During brushing, the brushing hand moves continuously either up-down or left-right. Therefore, we observe periodicity in accelerometer signals (see the left segment in the top figure). On the other hand, during flossing with string, there is a synchronized motion of both wrists (see the right segment in the bottom figure).

Figure 4 represents accelerometer, gyroscope, and orientation (roll, pitch, and yaw) signals during brushing with a manual toothbrush, followed by rinsing, and then flossing with a string. We make some key observations that guide our model development. First, we note that during brushing, the wrist is above the elbow, indicated by the lateral axis of accelerometer being positive. Second, we observe the repetitive motion of the wrist indicated by high motion in all three axes of the accelerometer. Third, during flossing with string, there is a synchronized motion of both wrists, indicated by the synchronized motion of each of the three orientation axes in both wrists. But, the magnitude of motion during flossing is low, which presents challenges in reliably distinguishing it from other daily behaviors. We now describe our approach to addressing these modeling challenges.

4.3 Overall Approach

We break down the overall problem of reliably detecting oral health behaviors from wrist-worn inertial sensors into several sub-problems for clarity. First, the signal processing unit needs the placement configuration of the wrist-worn device. Therefore, data needs to be transformed into the wrist coordinate system before further processing. Second, our goal is to detect rare OHB events from a continuous stream of sensor data. To significantly reduce the amount of sensor data that the final machine-learning model is applied on, we identify candidate windows that are most likely to contain OHB events. Third, after locating the candidate windows, we need to identify, select, and compute features to train a classifier. Finally, the utility of detecting these events depends

:wr



(a) Prevalent wrist orientations (b) Prevalent wrist positions (c) Kernel density function to show prevalent wrist positions and orientations

Fig. 5. (a) *Prevalent wrist orientations*; the value of a_p varies with the movements of rotating the forearm (b) *Prevalent wrist positions*; the value of a_l varies for different positions of wrist relative to elbow (c) *Kernel density function* to show prevalent wrist positions and orientations. The marginal distributions, at the top for perpendicular axis and at the right for the lateral axis, show the distribution of different measurements of the accelerometer sensor over a day (using $g = 9.8ms^{-2}$).

on correctly identifying the start and end of each event. Therefore, we need to identify the boundaries of the detected OHB event.

4.4 Virtual Orientation

As described in Section 4.1, in the free-living natural environment, and due to diversity in devices, mounting and wearing, configurations are unknown and can even dynamically change each time the device is taken off and put back on. Hence, we need to determine the orientation of the three axes with respect to the wrist using only sensor data, i.e., accelerometer and gyroscope traces.

The inertial sensor itself can be mounted in the device in $3! \times 2^3 = 6 \times 8 = 48$ possible configurations. This is because each of the axes can be matched to one of the three real-world dimensions (in $3! = 6$ ways). Next, any axis can be pointing upwards or downwards (in $2 \times 2 \times 2 = 8$ ways). But, for any wrist-worn device, the mounting of the inertial sensors is set at the time of design itself. Subsequently, the device can be worn in four different configurations (see Figure 3b). We can use a configuration file to denote this mounting, reducing the number of dynamic changes in orientation to four. Therefore, the problem of virtual orientation is about determining which of these four configurations best represents the data being collected. Further, we note that as gyroscopes and accelerometers are on the same chipset, the orientation of accelerometers also determines the orientation of the gyroscopes.

To develop and evaluate a solution to the virtual orientation problem (to determine which of the four configurations the sensor currently is in), we need labeled training data. To construct such a labeled dataset, we use the video capture that occurred twice daily (at the time of performing OHBs). We labeled a day's worth of data belonging to a particular configuration by observing the sensor wearing position on the wrist from videos. If we observed consistent configuration in both videos, then we labeled that day's data to belong to this

Algorithm 1: Pseudo code for finding correct configuration of a wristband

Input: $\vec{a} = \{a_s(t) = (a_x, a_y, a_z)\}$, D
Output: $c \in \{0, 1, 2, 3\}$
for *each* $c \in \{0, 1, 2, 3\}$ **do**
  $\vec{a_c} = \vec{a} * M_c$;
  $S_c = d(a_c, D)$;
end
return $c = \arg\min_c \{S_c\}$

particular configuration. In our data set, 77.08% (148 out of 192 days) of the data corresponds to Configuration 0 (see Figure 3). We label this as our base configuration. We create a database, D, with all the data corresponding to Configuration 0 (from 148 person days).

We use $\vec{a}$ to represent a time series of data points within one window of unknown configuration. We consider multiple window sizes (e.g., half an hour, 1 hour, 1.5 hours, 2 hours, 2.5 hours, etc.).

Our approach involves determining which of the four configurations make the data most similar to the base configuration upon its transformation via matrix multiplication. If we know the correct wearing configuration, $c \in \{0, 1, 2, 3\}$, then translating data from that configuration to the wrist coordinate system can be done via a matrix multiplication, i.e., $a_c(t) = a(t) * M_c$ for all t. For the four possible configurations, corresponding matrices (M_c) are as follows:

$$M_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; M_1 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; M_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}; M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

For example, for any accelerometer sample, (a_x, a_y, a_z), of Configuration 0, the corresponding value in the wrist coordinate system is $(a_l, a_p, a_v) = (a_x, a_y, a_z) * M_0 = (a_x, a_y, a_z)$.

Now, for a given window of accelerometer sensor data $\vec{a}$, if $c \in \{0, 1, 2, 3\}$ is the correct configuration, then $\vec{a_c} = \vec{a} * M_c$ represents data in the wrist coordinate system. We assume that similarity between $\vec{a_c}$ and D is greater than any other $\vec{a}_{c'} (c' \neq c)$. To find the correct c for a given $\vec{a}$ based on this assumption, we first create four transformations of $\vec{a}$, i.e., $\vec{a_0}$, $\vec{a_1}$, $\vec{a_2}$, and $\vec{a_3}$. We then compute the similarity index between each $\vec{a_c}$ and D. We need an efficient method to compute the similarity between $\vec{a_c}$ and D as there can be millions of data points in $\vec{a_c}$ and D.

To explore the potential for significant data reduction, we analyze the distribution of out dataset. Figure 5c shows the joint density of different positions and orientations of the wrist along the lateral and perpendicular axes. The marginal distributions, at the top for the perpendicular axis, and at the right for the lateral axis, show the distribution of different sensor measurements over a day. We observe two clusters in the joint density plot which represent the three most common orientations of the wrist, i.e., pronation, supination, and neutral. This guides us to develop two succinct representations of the data —probability distribution and principal components. We consider the following two corresponding indices of distance $d(\vec{a}, \vec{D})$.

(1) Distribution distance, $d_{dist}(\vec{a}, D)$: We compute probability distributions of the data points in $\vec{a}$ and D. Let P be the distribution of $\vec{a}$ and Q be the distribution of D. We then use earth moving distance between P and Q to find $d_{dist}(\vec{a}, D)$.
(2) Distance of principal components, $d_{PCA}(\vec{a}, D)$: We compute three major directions of data by computing principal component analysis (PCA). Let $\vec{pca}_a$ and $\vec{pca}_D$ be the vector of three major components of all

data points in $\vec{a}$ and D respectively. Then, we use the Euclidean distance between $\vec{pca}_a$ and $\vec{pca}_D$ to find $d_{PCA}(\vec{a}, D)$.

We compute distances using each of these two methods for all values of $\vec{a}$ and D for each configuration c. The window of data $\vec{a}$ is assigned to Configuration c that results in the minimum distance. Algorithm 1 describes the entire process.

4.5 Candidate Identification

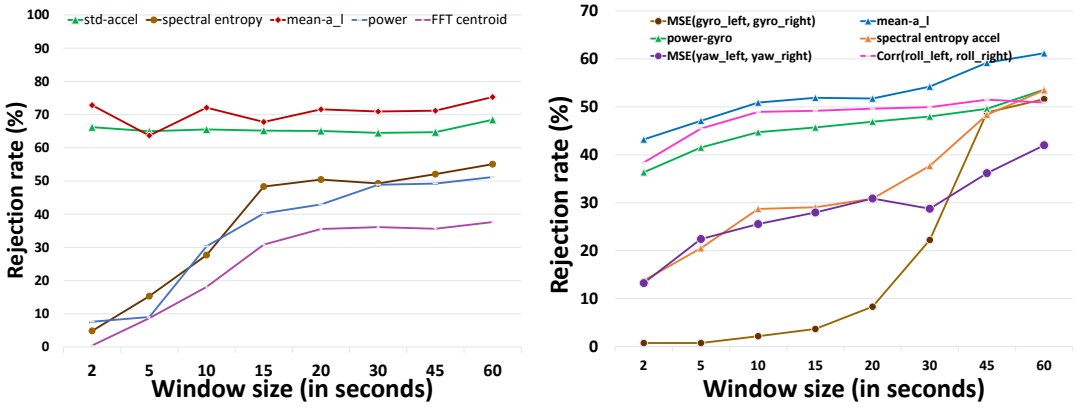
Our goal is to have a passive detection model that mines the continuous stream of sensor data to identify brushing and flossing events automatically. Each day, the sensor may be worn for an average of 16 hours (the awake portion of the day), but the events of interest (brushing or flossing), i.e., the positive class, may last only 4 minutes. Hence, almost all data represents the negative class. It is desirable to have a highly efficient model that can run on streaming data in real-time and rule out the majority of the data that do not contain the events of interest. This will reduce the amount of data on which a more complex model will need to be applied. We refer to these data segments as *candidate segments*. This staged detection approach can also help to reduce the false positive rate.

We explore two approaches for candidate identification—window-based and event-based. In the fixed window-based approach, we segment the time series into equal-sized windows and then compute some features to quickly detect whether they should be considered candidates. In the event-based approaches, we identify markers in the time series that may indicate a potential start of the event of interest. We observe that during brushing and flossing, the wrist position is usually higher than the elbow (see Figure 5b). We detect upward-based and downward-based wrist movements to identify and isolate segments of data that become candidates.

4.5.1 Toothbrushing candidate identification based on fixed windows. In the *fixed window-based approach*, we vary window sizes from 2 seconds to 60 seconds to find the most appropriate window size. For each window, we compute the time domain and frequency domain features to identify the most informative feature(s). Figure 6a shows the percentage of filtered windows per feature and per window size. We select the mean value of the lateral axis of the accelerometer as our primary filtering criteria because it produces the highest rejection rate. The next best feature is the standard deviation of accelerometer magnitude. We observe that both of these features are stable across all window sizes and consistently provide a rejection rate of higher than 60%. For window sizes of 15 seconds or higher, most of the features have a stable rejection rate. A smaller window size is preferred so as to more accurately isolate pause events within brushing episodes. Hence, we use 15 seconds as our window size. We also consider using two features together to see if they provide a substantial gain in the rejection rate. After fixing the window size (to 15 seconds), using the top two features increases the rejection rate from below 70% (for the top feature) to 75.6% for a window size of 15 seconds.

4.5.2 Toothbrushing candidate identification based on events. We observe that the wrist position is usually higher than the elbow during brushing and flossing. In PuffMarker [Saleheen et al. 2015], hand-to-mouth gestures were detected as the candidate segments to detect smoking puffs. While taking a smoking puff, the wrist usually stays stationary when close to the mouth. Therefore, the magnitude of the gyroscope was used to locate stationary periods in the time series that includes hand-at-mouth events. But, during brushing, the hand moves continuously. Hence, this method cannot be directly used to find candidates for OHBs.

We use detected upward- and downward-based wrist movements to generate segments, depicted in Figure 5b. First, we find a threshold TH_l for the lateral axis of accelerometer that filters all samples below this threshold. We optimize threshold TH_l such that the amount of filtered data is maximized while ensuring that all data from the positive class are preserved. We find the optimal value to be 0.24 g (using $g = 9.8ms^2$). This method filters 81% of the data. It, however, creates some temporal clusters in the data. If the time difference between consecutive clusters of retained samples is less than 1 second, then they are merged to create candidate segments. These small



(a) Rejection rate for brushing

(b) Rejection rate for flossing

Fig. 6. Rejection rate when using different features as window size is varied for (a) manual brushing and (b) flossing with string

gaps can be created due to jerks or spikes. This process creates several small candidate segments because of the yawn or upper body touch or other frequent transient behaviors. To prune such segments, we apply another filtering based on the time duration of the candidate segments. We find values of min_{dur} and max_{dur} such that it includes all the positive events but filters as many false candidates as possible. We find that $min_{dur} = 11$ seconds and $max_{dur} = 2.5$ minutes work best for our dataset. It filters out 10% of the additional data. Altogether, this method rejects 91% of the data and results in an average of 100 candidate segments per day. Comparatively, the window-based approach with a 75% rejection rate can produce approximately 1,000 candidate segments in 16 hours of sensor wearing.

4.5.3 Flossing candidate identification based on fixed size window. Similar to our candidate identification approach for brushing, as shown in Figure 6b, we vary window sizes from 2 seconds to 60 seconds. For each window, we compute various correlation-based features as the orientations of both wrists remain similar during flossing with string. Figure 6b represents the percentage of filtered windows per feature and per window size. We observe that for the top feature (mean value of lateral axis of the accelerometer), there is an increase of 10% in the rejection rate, but for the second and third performing features, the increase is only marginal. As flossing events can last less than a minute and involve pauses, we use 15 seconds as the window size. We combine the top two features in our filtering criteria that together provide a rejection rate of 62.3% (for a 15-second window size).

4.5.4 Flossing candidate identification based on event. Figure 1b shows several images during flossing with string. We observe that during flossing, both wrists are in the upward direction. Hence, an approach similar to brushing candidate identification can work for flossing candidate identification. We can achieve further pruning for flossing because flossing usually involves both wrists. We use the same filtering method that is used for brushing to identify candidate segments for flossing, but here we compute two sets of candidates from each of the wrists. We retain only those segments that have overlap in data from both wrists. Using the same duration threshold as in brushing candidate identification, we filter out any candidate segment whose overlapped duration is outside of the range. This method rejects 95% of the data and results in an average of 70 candidate segments per day. This represents a significant (32%) increase in rejection rate compared to the fixed window-based approach.

Table 2. Summary of features extracted from each segment of the data

Time-domain (td)	**Frequency-domain (fd)**	**Multi-sensor fusion orientation (ori)**	**Cross-wrist (only for flossing)**
mean	Maxima	Roll	corr(roll-left, roll-right)
median	Energy	Pitch	corr(pitch-left, pitch-right)
standard deviation	Spectral centroid	Yaw	corr(yaw-left, yaw-right)
quartile deviation	Spectral Flux		root Mean Square Error (rMSE)
skewness	Spectral roll-off		
kurtosis			
zero-crossing			
power			

4.6 Feature Computation and Selection for Candidate Classification

After identifying the candidates, we compute several time domain, frequency domain, multi-sensor fusion, and cross-wrist features from accelerometer and gyroscope data in each window. Table 2 summarizes all the computed features. For *time domain features*, we compute the mean, median, standard deviation, quartile deviation, skew, kurtosis, and zero crossing rate of three axes of accelerometer and gyroscope. For *frequency domain features,* Fourier transformation is applied on the window of data before calculating common frequency-domain features [Dargie 2009]. For *wrist orientation features*, we compute roll, pitch, and yaw that provide information about the orientation of the wrist with respect to gravity.

To obtain robust measurement of orientation features from noisy inertial sensor data, we perform the following processing steps [Jain et al. 2015; Min and Jeung 2015]. Accelerometer, $a(t)$, gives a good indication of orientation in static conditions. Gyroscope, $\omega(t)$, provides a good indication of tilt in dynamic conditions, but drifts in the long term. The value of $a(t)$ is noisy, but over longer intervals is useful, as it is more robust to drift [Jain et al. 2015; Min and Jeung 2015]. By passing the accelerometer signals through a low-pass filter, passing the gyroscope signals through a high-pass filter, and combining the resultant signals, we compute a final rate function. The idea behind *complementary filter* is to take slow moving signals from $a(t)$, fast moving signals from $\omega(t)$, and combine them. This method combines the strengths of both sensor signal streams. From Figure 4, we observe that change in the orientation of both wrists is similar, so we compute correlation and root mean square error of orientation of left wrist and orientation of right wrist. We compute these *cross-wrist features* only for flossing because flossing with string requires both wrists.

In total, we obtain more than 100 features. But, to avoid overfitting (as there are only 160 brushing and 137 flossing events), we use selected features for modeling. The idea behind feature selection is to remove non-informative features. Our goal is to find a subset of features that are a) mutually not correlated but b) highly correlated to the OHBs. In this work, we used the Correlation-based Feature Selection (CFS) [Hall 1999] to select a subset of the features (25) as in other detection based works [Rahman et al. 2016; Sarker et al. 2014]. CFS selects features that are mutually uncorrelated but highly indicative of the OHB classes. We describe feature selection further in Section 6.3.

4.7 Model Selection and Training

We consider the following models for both toothbrushing and flossing classification. We use grid search to optimize the hyper-parameters in each model.

- **Naive Bayes classifier (NB)**: Naive Bayesian networks are very simple Bayesian networks which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent [Kotsiantis et al. 2007].
- **Random forest classifier (RF)**: Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifier depends on the strength of the individual trees in the forest and the correlation between them [Breiman 2001]. We use Random forests with 100 trees and 1,000 trees.
- **Ensemble method (Ens)**: Ensemble classifiers [Dietterich 2000; Rokach 2010] consist of a set of many individual classifiers (called base-learners) where those decisions are combined to output a single class label. Ensemble learning helps improve machine learning results by combining several models. This approach allows for the production of better predictive performance compared to a single model. Several different methods are proposed to address data imbalance issues by using ensemble methods [Estabrooks et al. 2004; He and Garcia 2008]. Ensemble models also tend to generalize better, which makes this approach easy to handle. In our experiments, we use a combination of Decision tree, K-nearest neighbors (KNN), and Naive Bayes [Caruana et al. 2004].
- **Ada-Boosting (AB)**: AdaBoost, short for *Adaptive Boosting*, is the first practical boosting algorithm proposed by [Freund et al. 1996]. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. This is also an ensemble method that learns models on subsets of the training data and boosts the weights of misclassified instances, which allows models to focus on those for improving classification performance.

5 IMPLEMENTATION

Figure 7 presents an overview of all the data processing steps that include screening, cleaning, finding window size, selecting candidates, computing the features, and training the machine learning model to detect oral health behaviors.

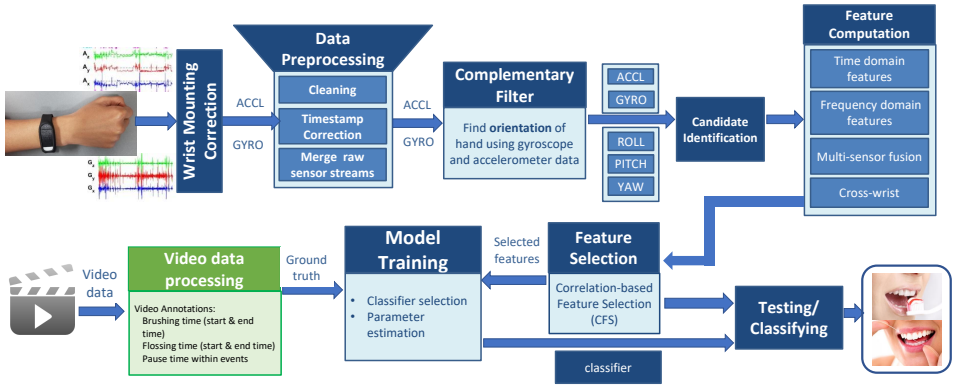
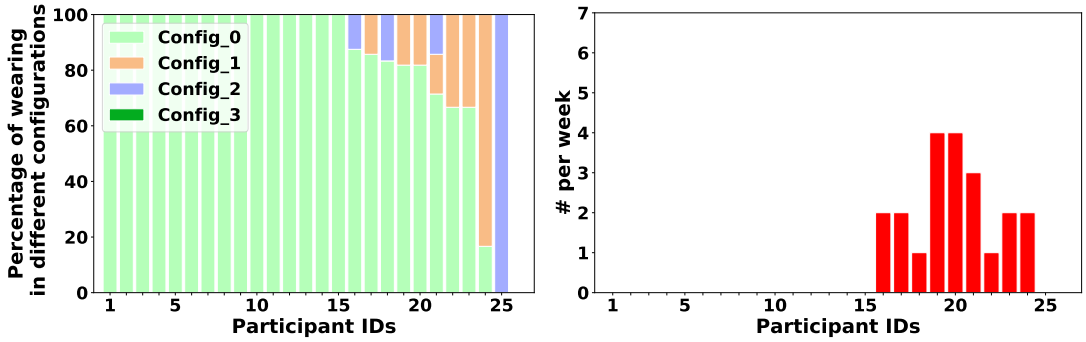


Fig. 7. Data processing stages for training and testing models for brushing and flossing detection.



(a) Use of different configurations by participants (b) Frequency of configuration changes by participants

Fig. 8. Prevalence of sensor configurations among participants and the frequency of configuration changes.

Briefly stated, we first determine the sensor mounting on the wrist. Second, we remove data segments when the sensor was not worn by the participant. Third, we impute intermittent data when the gap is less than 0.25 seconds using methods presented in [Rahman et al. 2014; Saleheen et al. 2015]. Fourth, we use *complementary filter* to compute roll, pitch and yaw by combining accelerometer and gyroscope data. Fifth, we locate and mark candidate windows. Finally, we compute features from those candidates. For classification, Ada-boost produces the best results.

6 MODEL EVALUATION

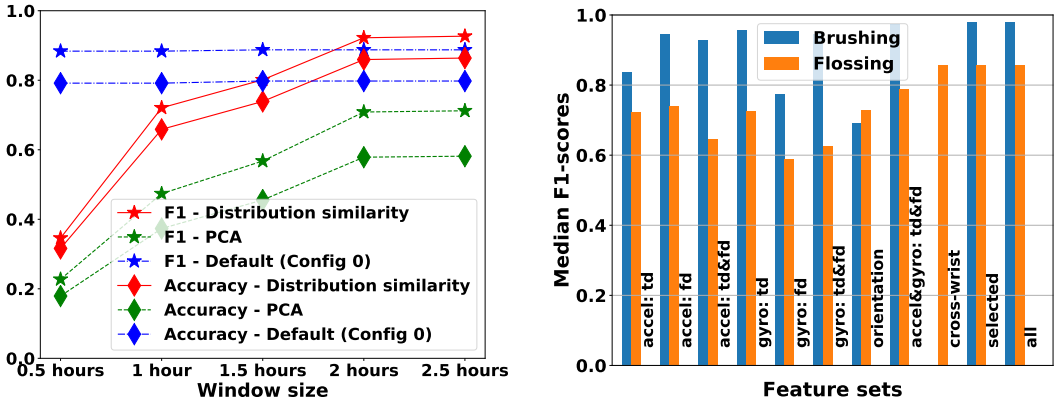
We now present results from our evaluation. We start with comparing the performance of identifying wrist configuration correctly for the distribution-based method and PCA-based method. Then we observe the performance of the selected feature set for detecting OHB events. We use the selected feature set for further experiment. We compare the performance of detecting brushing and flossing for different models choices. We evaluate the impact of using window-based or event-based approaches to candidate identification, as they impact which segments of data the machine learning models are applied to. Finally, we analyze the error in duration and start/end times of the detected brushing and flossing events.

6.1 Performance of Virtual Orientation

To identify the virtual orientation, we determine the window size and the method that provides best performance (see Section 4.4). We consider three methods for virtual orientation — always use a default configuration, use PCA, and use distribution similarity for virtual orientation. We create a labeled data for this experiment by dividing the entire dataset in specific window sizes. All data for a specific day are labeled as belonging to the configuration that was observed from two videorecordings for that day. Each of the three methods are applied to each window of data and the resulting configuration is compared with the label assigned to that window.

We use 192 days of sensor data from 25 participants. Figure 8a presents the percentage of wristband wearing in different configurations. We observe that 15 (out of 25) participants always wear the sensor in Configuration 0 (see Figure 8b). Nine participants switch configurations across days, and one participant always wears it in Configuration 2. Observing this pattern, we make Configuration 0 as the default configuration.

We consider window sizes of half an hour, 1 hour, 1.5 hours, 2 hours, and 2.5 hours. From Figure 9a, we observe that the performance saturates after a window size of 2 hours. Therefore, we select our window size as 2 hours. As expected, the choice of window size has no effect on the performance of the default configuration. The



(a) F1-score and accuracy for virtual orientation

(b) F1-scores for brushing and flossing detection using different feature sets using LOSOCV evaluation

Fig. 9. (a) Impact of window size on F1 and accuracy for identifying the correct configuration, and (b) Median F1-scores for detecting brushing and flossing using selected features

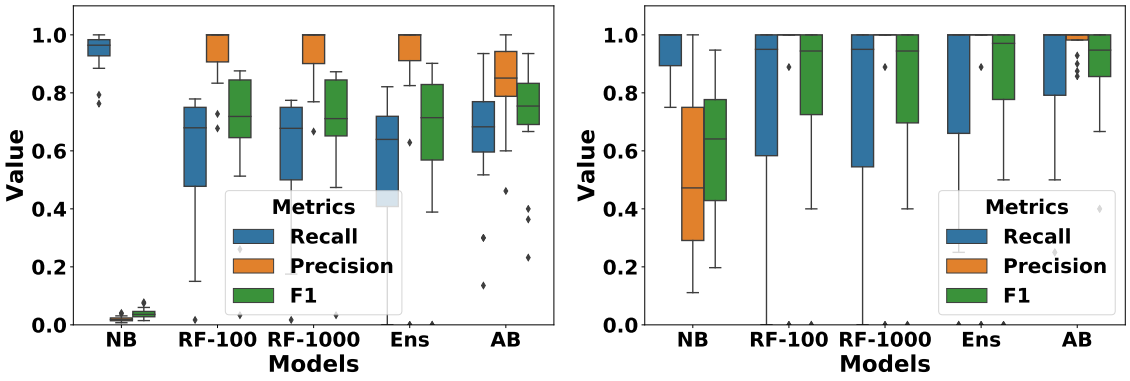
distribution-based method produces better results (an accuracy of 86% and F1-score of 93%) than both the default (an accuracy of 79% and F1-score of 88%) and PCA-based method (an accuracy of 58% and F1-score of 71%). For real-life deployment, one can start with the default configuration and then use the distribution based method after two hours of data has been collected to personalize the virtual orientation to each user.

6.2 Metric for Evaluating Performance of Classifiers

For classifaction performance evaluation, we use a leave-one-subject-out cross-validation (LOSOCV) experiment where we exclude a single user's data for testing and use the remaining data for training. We repeat this procedure for all the users. We present results in box plots [Stisen et al. 2015] and report median values. We note that in our dataset (similar to real-life usage), only 0.1% of the data are positive instances and rest 99.9% of the data are negative instances. If accuracy alone is used to measure the classification performance, then a simple model that classifies all testing samples into the negative class, will produce an excellent accuracy of 99.9%. But, its recall and precision will be zero and F1-score (for detecting the positive class) will be undefined. Since our interest is in reliably detecting the positive class, we use recall, precision, and F1-score to measure the performance of our classifiers. In addition, we report false positives detected per day to provide a sense of the model's performance in daylong wearing.

6.3 Feature Selection

We evaluate the discriminatory power of different feature sets. Figure 5 shows different feature sets and their discriminatory power. We create a set of time-domain features, a set of frequency-domain features, and finally a set of both time-domain and frequency-domain features from the accelerometer data. We perform the same selection from the gyroscope data. Next, we create a set of orientation related features. We create another set by combining both accelerometer and gyroscope features. We also create another set using cross-wrist features that is used only for flossing detection. Finally, one feature set is generated using the feature selection algorithm described in 4.6. To detect OHBs, the selected feature set (consisting of 25 features) performs almost the same as using all the features. To understand the performance of the proposed model, independent of the error in



(a) Performance of brushing detection with a (15 seconds) window-based approach for candidate identification.

(b) Performance of brushing detection with an event-based approach for candidate identification.

Fig. 10. Box-plots of leave-one-subject-out cross validation results for brushing detection. NB: Gaussian Naive Bayes, RF-100: Random Forest with 100 trees, RF-1000: Random Forest with 1000 trees, Ens: Ensemble method, and AB: Ada-Boosting

detecting correct orientation, this evaluation and the rest of the experiments are done using the proper virtual orientation of the wrist.

6.4 Performance of Brushing Detection

We evaluate both window-based and event-based approaches to candidate identification (see Section 4.5) for their impact on classification performance. For both experiments, Gaussian Naive Bayes, Random Forest with 100 trees and 1000 trees, Ensemble methods, and Ada-boost classifiers are applied to detect brushing behaviors.

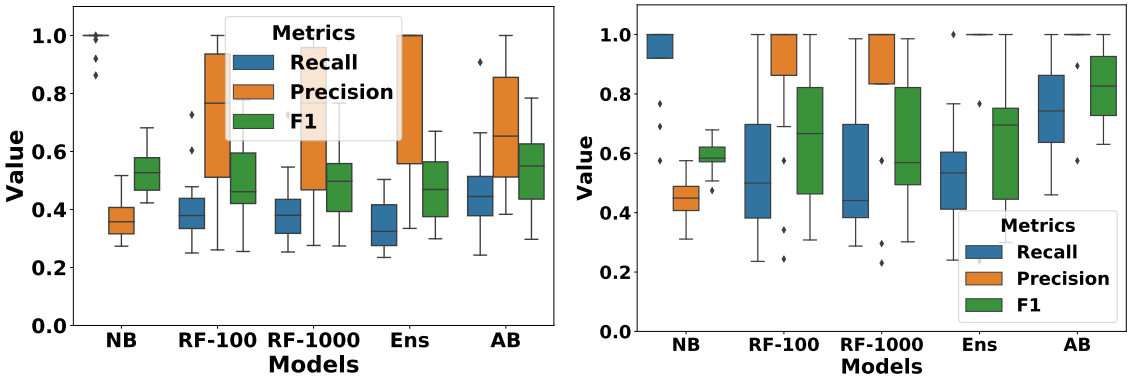
6.4.1 Window-based approach. As described in Section 4.5, we use a window size of 15 seconds. Recall, precision, and F1 scores appear in Figure 10a. For the Naive Bayes model, median precision rates are lower than other models, but the recall rate is the highest. Ensemble method, Random Forest with 100 trees and 1000 trees produce a high median precision rate of 100%, but recall rates of only 64.74%, 67.15%, and 68.27% respectively. Ada-boost's median recall rate is 70.44%, and a precision rate of 85.15%. The Ensemble method and Random forest with 1000 trees produce a median F1 scores of respectively 71.43% and 71.86%, but the highest median F1-score, i.e., 75.58%, is produced by Ada-boost.

6.4.2 Event-based approach: Recall, precision, and F1 scores when using event-based approach for candidate generation are shown in Figure 10b. The performance of all the models are higher than that for a window-based approach. The Ada-boost model provides the best F1 score. It has a median recall of 100%, a precision of 100%, and an F1 score of 95%, which is 19.42% higher than that for the window-based approach. With the Ada-boost model, we get about one false positive every nine days (i.e., 16 false positives on 151 person days of data).

6.5 Performance of Flossing Detection

We follow a similar approach for evaluating the detection of flossing as for brushing detection.

6.5.1 Window-based approach: Recall, precision, and F1 scores appear in Figure 11a. We obtain the best F1 score (a median F1 score of 55.02%) for Ada-boost. It has a median precision of 65.30%, and a median recall rate of 44.50%.



(a) Performance of flossing detection with a (15 seconds) window-based approach for candidate identification.

(b) Performance of flossing detection with an event-based approach for candidate identification.

Fig. 11. Boxplot of leave-one-subject-out cross validation results for flossing detection. NB: Gaussian Naive Bayes, RF-100: Random Forest with 100 trees, RF-1000: Random Forest with 1000 trees, Ens: Ensemble method, and AB: Ada-Boosting

6.5.2 Event-based approach: Recall, precision, and F1 scores are shown in Figure 11b. Similar to brushing, we obtain the best median F1 score with the Ada-boost model. Its median recall is 75%, precision is 100%, and the F1 score is 82.65%, which is 27.63% higher than that for a window-based approach. With the Ada-boost model, we obtain about one false positive every thirty-two days (i.e., 6 false positives on 192 person days of data).

We observe that our detection accuracy for flossing is lower than that for brushing. One reason for a lower recall rate in detecting flossing is due to low and infrequent movement of hands (captured in inertial sensor data) during flossing. Improving the recall rate for flossing detection remains a future work.

6.6 Performance on detecting duration and start/end times

In addition to reliably detecting brushing and flossing events, it is also desirable to accurately detect the duration of these events (due to their predictive nature in dental disease outcomes). Further, as these events are detected from a continuous time series of sensor data, a single event can be split into multiple windows due to pauses. Portions of the actual event can also be missed or overestimated. We would like the detected event to closely match the actual event in both the start and end times. We first develop some metrics to measure the accuracy of detecting start/end times and duration before reporting the performance of our models.

Event: An event consists of a start time and an end time. We define a pause event as a tuple of start time and end time ($E_p = (t_s, t_e)$). Pause events can be either inter-event pause or intra-event pause. Since we are not dealing with inter-event pauses, we use pauses to refer to intra-event pauses. Each oral hygiene event consists of a start time, an end time, and a list of pause events. For example, $(E_b = (t_s, t_e, E_p))$ is a brushing event with a list of pause events, where $t_s < e.t_s, e.t_e < t_e$ for all $e \in E_p$ and $(E_f = (t_s, t_e, E_p))$ is a similarly defined flossing event .

Duration of an event: The duration of an event $E = (t_s, t_e, E_p)$ (after removing pauses) is defined as follows.

$$d(E) = (E.t_e - E.t_s) - \sum_{E_p \in E.E_p} (E_p.t_e - E_p.t_s)$$

Error in event localization: We divide the error by the duration of the actual event for normalization. Let E_{actual} be the actual event. For a detected event, $E_{detected}$, we use two error measurements:

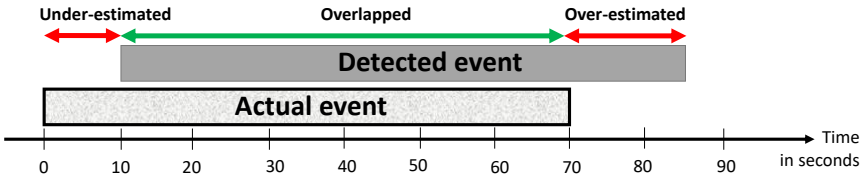
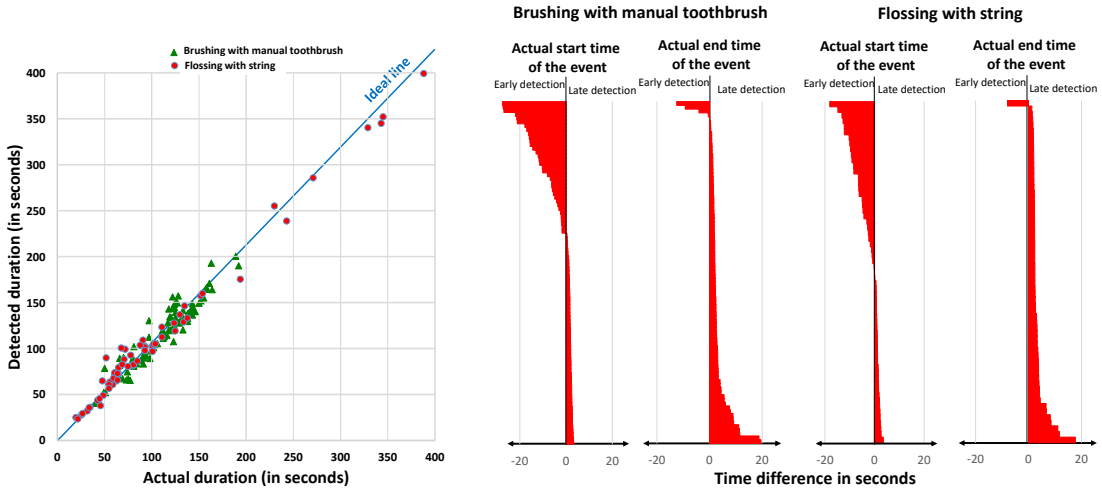


Fig. 12. Definition and an example of duration error ($err_{dur} = \frac{|70-75|}{70} = \frac{5}{70}$) and Localization error ($err_{boundary} = \frac{|10+15|}{70} = \frac{25}{70}$)



(a) Scatter plot of actual duration vs. detected duration, err_{dur} = 7.2% for brushing and 6.2% for flossing.

(b) Error in start/end times, $err_{boundary}$ = 4.1% for brushing and $err_{dur} = err_{boundary}$ = 3.5% for flossing.

Fig. 13. Performance for detecting the duration and start/end times for brushing and flossing.

- *Duration error* is defined as the difference between actual duration and detected duration, i.e.,

$$err_{dur} = \frac{|d(E_{actual}) - d(E_{detected})|}{d(E_{actual})}$$

- *Average localization error* of the event measures error in locating the boundary. Error can happen in both boundaries due to either over-estimation or under-estimation. We define average localization error as

$$err_{boundary} = \frac{AVG\{|E_{actual}.t_s - E_{detected}.t_s| + |E_{actual}.t_e - E_{detected}.t_e|\}}{d(E_{actual})}$$

Figure 12 shows a scenario where duration error is low but localization error is high.

For each brushing and flossing event detected by the best model, we evaluate the error in the duration of the detected event with that of the corresponding actual event (from video data). If a detected event does not have any temporal overlap with the actual event, it is not considered to be a true recall and is excluded from this analysis. We use regression analyses to analyze the errors in duration.

Figure 13a depicts regression analyses between the actual durations and the corresponding durations of the detected events (for both brushing and flossing). The Pearson correlation coefficient value r is 0.95 for brushing and 0.98 for flossing. Mean square error for brushing is 11.07 seconds and for flossing, it is 8.9 seconds. As a percentage of the actual event duration, these represent duration errors of 7.2% and 6.5% respectively.

Next, we analyze the temporal alignment of the detected event with the actual event. Figure 13b shows the results for brushing and flossing respectively. We observe that the error in alignment is small for both events. The error in accurately locating the start and end times are 4.1% for brushing and 3.5% for flossing.

7 LIMITATIONS AND FUTURE WORKS

The presented mORAL model achieves a median recall rate of 100% with one false positive every nine days. For flossing, the median recall rate is lower at 75%, but the false positive rate is also lower at one every 37 days. In addition to low recall rate for flossing, this work has several other limitations that open up numerous opportunities for future works.

First, our model does not include oral rinsing behavior detection, which is an important oral health behavior. Especially when combined with a model for detecting eating, it can present interesting intervention opportunities. Modeling the transition in the sequence of brushing, rinsing, and flossing behavior can potentially be used to improve the detection of each of these activities. Second, our model can detect flossing with string, but not with picks. Using deep learning models, especially with a larger labeled dataset, can potentially improve detection performance further.

Third, a significant limitation in this work is the use of two wristbands. We used two wristbands to understand the value of having the wristbands in both hands. We note that detection of brushing uses only the data from dominant hand. Therefore, detection of brushing should work as long as the user is wearing the sensor in their dominant hand. But, the model needs to be adapted to detect flossing from sensor data collected on only one wrist.

Fourth, detecting brushing with the SmartBrush from wristband is another future research direction. Even though smart toothbrushes detect the brushing event due to the user pressing a button to start/stop the device, a wrist-based model can be used to assign the brushing event to a specific user, e.g., when a brush handle is shared in a family.

Fifth, even though significant work has been done in detecting the brushing pressure and detecting the tooth surface or quadrant being brushed, using smart or instrumented toothbrushes, our work opens up opportunities to develop these capabilities for brushing with manual toothbrushes, especially due to low boundary location error with our model. Doing so will benefit a large population that still uses manual toothbrushes.

Finally, additional work is needed to adapt our model for detecting the brushing and flossing events in real-time. Although our model is computationally efficient, the virtual orientation process may require a couple hours of data. Developing a quicker method for virtual orientation can help not only with our model, but also in models developed for detecting other behaviors such as eating and smoking that involve hand-to-mouth gestures.

8 CONCLUSIONS

mHealth research is rapidly growing to incorporate detection, prediction, and just-in-time intervention for a diversity of markers of health and well-being. Our work extends this paradigm into the oral health domain, an oft-underserved realm of biomedicine and public health. It opens doors for extending several benefits only available to those with access to smart toothbrushes to a broader population still using manual toothbrushes. It also opens up new opportunities for designing sensor-triggered interventions for improving oral health behaviors. For example, incorporating the passive detection of OHB's into a context-based approach can enhance our understanding of how OHB biomarkers are associated with other biomarkers such as stress, semantic location,

tobacco, alcohol, or other substance use, and may provide substantial future insight to these behaviors and lead to new supporting protective interventions.

In addition to providing a solution for monitoring oral health behaviors, our work also provides a direction for the computing community when developing models to reliably detect rare daily events from dense time series of data collected in everyday life. For example, it shows the promise of the event-based approach for identifying candidate windows before applying the machine learning models. It further indicates that identifying discerning characteristics of the target event and translating them into efficient data models results into a more accurate overall model.

9 ACKNOWLEDGEMENTS

We thank the anonymous reviewers and Mithun Saha from University of Memphis for editorial help. We also thank Dr. Benjamin Marlin from UMass Amherst for his advice on modeling virtual orientation. Research reported here was supported by the National Institutes of Health (NIH) under award R01DE024244 by the National Institute of Dental and Craniofacial Research and awards R01CA224537, R01MD010362, R01CA190329, R24EB025845, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative. It was also supported by the National Science Foundation (NSF) under awards IIS-1722646, ACI-1640813, and CNS-1823221. The authors wish to acknowledge the material and technical support provided by Hansjoerg Reick and Ingo Vetter from Oral-B/Procter & Gamble. The opinions expressed in this article are authors' own and do not reflect the view of the NIH, NSF, or Oral-B.

REFERENCES

Regina M Benjamin. 2010. Oral health: the silent epidemic. *Public health reports* 125, 2 (2010), 158–159.

Lal Gamze Bozgeyikli, Evren Can Bozgeyikli, and Andrew Raij. [n. d.]. Keep Brushing! Developing Healthy Oral Hygiene Habits in Young Children with an Interactive Toothbrush. ([n. d.]).

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

Barbara Bruno, Fulvio Mastrogiovanni, and Antonio Sgorbissa. 2014. A public domain dataset for ADL recognition using wrist-placed accelerometers.. In *RO-MAN*. 738–743.

Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. 2012. Human motion modelling and recognition: A computational approach. In *2012 IEEE International Conference on Automation Science and Engineering (CASE 2012)*. IEEE, 156–161.

Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. 2013. Analysis of human behavior recognition algorithms based on acceleration data. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 1602–1607.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 18.

CDC et al. 2010. Oral health: Preventing cavities, gum disease, tooth loss, and oral cancers. At a glance 2011. *Atlanta: Division of Oral Health, National Center for Chronic Disease Prevention and Health Promotion* (2010).

Barbara Lesley Chadwick, Deborah White, Deborah Lader, and Nigel Pitts. 2011. Preventive behaviour and risks to oral health-a report from the Adult Dental Health Survey 2009. (2011).

Yu-Chen Chang, Jin-Ling Lo, Chao-Ju Huang, Nan-Yi Hsu, Hao-Hua Chu, Hsin-Yen Wang, Pei-Yu Chi, and Ya-Lin Hsieh. 2008. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 363–372.

CHIIZ [n. d.]. Sonic-powered automatic toothbrush. https://bfbe9e.kckb.st.

Waltenegus Dargie. 2009. Analysis of Time and Frequency Domain Features of Accelerometer Measurements.. In *ICCCN*, Vol. 9. 1–6.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

Katherine Ellis, Jacqueline Kerr, Suneeta Godbole, Gert Lanckriet, David Wing, and Simon Marshall. 2014. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological measurement* 35, 11 (2014), 2191.

Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence* 20, 1 (2004), 18–36.

Yoav Freund, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *Icml*, Vol. 96. Citeseer, 148–156.

E Grossman, M Cronin, W Dembling, and H Proskin. 1996. A comparative clinical study of extrinsic tooth stain removal with two electric toothbrushes [Braun D7 and D9] and a manual brush. *American journal of dentistry* 9 (1996), S25–9.

Taku Hachisu and Hiroyuki Kajimoto. 2012. Augmentation of toothbrush by modulating sounds resulting from brushing. In *Advances in Computer Entertainment*. Springer, 31–43.

Taku Hachisu and Hiroyuki Kajimoto. 2015. Modulating tooth brushing sounds to affect user impressions. *International Journal of Arts and Technology* 8, 4 (2015), 307–324.

Mark Andrew Hall. 1999. Correlation-based feature selection for machine learning. (1999).

Haibo He and Edwardo A Garcia. 2008. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering* 9 (2008), 1263–1284.

Timothy Hnat, Syed Monowar Hossain, Nasir Ali, Simona Carini, Tyson Condie, Ida Sim, Mani B. Srivastava, and Santosh Kumar. 2017. mCerebrum and Cerebral Cortex: A Real-time Collection, Analytic, and Intervention Platform for High-frequency Mobile Sensor Data. In *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4-8, 2017*. http://knowledge.amia.org/65881-amia-1.3897810/t006-1.3900376/f006-1.3900377/2732068-1.3900423/2727701-1.3900420

Syed Monowar Hossain, Timothy Hnat, Nazir Saleheen, Nusrat Jahan Nasrin, Joseph Noor, Bo-Jhang Ho, Tyson Condie, Mani Srivastava, and Santosh Kumar. 2017. mCerebrum: A Mobile Sensing Software Platform for Development and Validation of Digital Biomarkers and Interventions. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 7.

Peng-Ju Hsieh, Yen-Liang Lin, Yu-Hsiu Chen, and Winston Hsu. 2016. Egocentric activity recognition by leveraging multiple mid-level representations. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 1–6.

Hua Huang and Shan Lin. 2016. Toothbrushing monitoring using wrist watch. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 202–215.

Shubham Jain, Carlo Borgiattino, Yanzhi Ren, Marco Gruteser, Yingying Chen, and Carla Fabiana Chiasserini. 2015. Lookup: Enabling pedestrian safety services via shoe sensing. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 257–271.

Kyeong-Seop Kim, Tae-Ho Yoon, Jeong-Whan Lee, and Dong-Jun Kim. 2009. Interactive toothbrushing education by a smart toothbrush system via 3D visualization. *Computer methods and programs in biomedicine* 96, 2 (2009), 125–132.

Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. 2015. Evaluating tooth brushing performance with smartphone sound data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 109–120.

Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. 2016. Toothbrushing performance evaluation using smartphone audio based on hybrid HMM-recognition/SVM-regression model. *Journal of Information Processing* 24, 2 (2016), 302–313.

Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* 160 (2007), 3–24.

Young-Jae Lee, Pil-Jae Lee, Kyeong-Seop Kim, Wonse Park, Kee-Deog Kim, Dosik Hwang, and Jeong-Whan Lee. 2012. Toothbrushing region detection using three-axis accelerometer and magnetic sensor. *IEEE Transactions on Biomedical Engineering* 59, 3 (2012), 872–881.

Cheng-Yuan Li, Yen-Chang Chen, Wei-Ju Chen, Polly Huang, and Hao-hua Chu. 2013. Sensor-embedded teeth for oral activity recognition. In *Proceedings of the 2013 international symposium on wearable computers*. ACM, 41–44.

Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. 2013. Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise* 45, 11 (2013), 2193.

Marco Marcon, Augusto Sarti, and Stefano Tubaro. 2016. Toothbrush motion analysis to help children learn proper tooth brushing. *Computer Vision and Image Understanding* 148 (2016), 34–45.

Hyung Gi Min and Eun Tae Jeung. 2015. Complementary filter design for angle estimation using mems accelerometer and gyroscope. *Department of Control and Instrumentation, Changwon National University, Changwon, Korea* (2015), 641–773.

Tatsuo Nakajima, Vili Lehdonvirta, Eiji Tokunaga, Masaaki Ayabe, Hiroaki Kimura, and Yohei Okuda. 2007. Lifestyle ubiquitous gaming: making daily lives more plesurable. In *Embedded and Real-Time Computing Systems and Applications, 2007. RTCSA 2007. 13th IEEE International Conference on*. IEEE, 257–266.

Zhenchao Ouyang, Jingfeng Hu, Jianwei Niu, and Zhiping Qi. 2017. An Asymmetrical Acoustic Field Detection System for Daily Tooth Brushing Monitoring. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 1–6.

Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 149–161.

Toby G Pavey, Nicholas D Gilson, Sjaan R Gomersall, Bronwyn Clark, and Stewart G Trost. 2017. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of science and medicine in sport* 20, 1 (2017), 75–80.

Md Mahbubur Rahman, Rummana Bari, Amin Ahsan Ali, Moushumi Sharmin, Andrew Raij, Karen Hovsepian, Syed Monowar Hossain, Emre Ertin, Ashley Kennedy, David H Epstein, et al. 2014. Are we there yet?: Feasibility of continuous stress assessment via wireless physiological sensors. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM,

479–488.

Tauhidur Rahman, Mary Czerwinski, Ran Gilad-Bachrach, and Paul Johns. 2016. Predicting about-to-eat moments for just-in-time eating intervention. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 141–150.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33, 1-2 (2010), 1–39.

Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 2015. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 999–1010.

Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 909–920.

Vivek Shetty, John Yamamoto, and Kenneth Yale. 2018. Re-architecting oral healthcare for the 21st century. *Journal of dentistry* 74 (2018), S10–S14.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 127–140.

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, 1029–1040.

Adrian Yap. 2017. Oral health equals total health: A brief review. *Journal of Dentistry Indonesia* 24, 2 (2017), 59–62.

Takuma Yoshitani, Masa Ogata, and Koji Yatani. 2016. LumiO: a plaque-aware toothbrush. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 605–615.

Shaoyan Zhang, Alex V Rowlands, Peter Murray, Tina L Hurst, et al. 2012. Physical activity classification using the GENEA wrist-worn accelerometer. (2012).

A APPENDIX

Table 3. Video annotation protocol for identifying OHB events and labeling their start and end times

Label	Meaning	Description and event characteristics
bn_st	brushing start time with manual brush	The first video frame when the participant's hand is close to the mouth and brushing begins
bn_ed	brushing end time with manual brush	When the participant stops brushing and wrist holding the manual brush is going down from the mouth
bo_st	brushing start time with Smart-Brush	Similar to the *bn_st* event, but for SmartBrush
bo_ed	brushing end time with SmartBrush	Similar to the *bn_ed* event, but for SmartBrush
fs_st	flossing start time with string	When the participant actively starts flossing with string (preparation time before flossing such as string winding around fingers) is excluded
fs_ed	flossing end time with string	When both of the participant's wrists move down from the mouth with string floss without subsequent resumption of flossing
fp_st	flossing start time with picks	When the participant flosses with a pick, and one wrist with pick moves up to the mouth and flossing begins
fp_ed	flossing end time with picks	When participant ends flossing with pick, and wrist moves down from the mouth
p_st	pause start	For brushing, pause begins when the participant temporarily stops brushing (e.g., to spit out the accumulated toothpaste or to rinse) and moves the wrist away from the mouth. For string flossing, pause begins when both wrists move away from the mouth. For pick flossing, pause begins when the wrist holding the pick moves away from the mouth
p_ed	pause end	Following *p_st*, pause ends for brushing when the brushing wrist goes up to the mouth. For string flossing, pause ends when both wrists go up to resume flossing. For pick flossing, pause ends when the wrist holding the pick moves back up to the mouth to resume flossing
pv	video pause time	A few participants paused the video when they paused brushing or flossing (e.g., to spit out the accumulated toothpaste or saliva from mouth), or sometimes between the brushing and flossing events. During these times, we missed all ground truth video. We annotate those events as video pause time.
ori_left	orientation of left wrist	We observe the device orientation on the left wrist in video and label the configuration as one of $\{1, 2, 3, 4\}$
ori_right	orientation of right wrist	We observe the device orientation on the right wrist in video and label the configuration as one of $\{1, 2, 3, 4\}$
bn_wt	brushing wrist	The brushing wrist is marked as left or right
fl_wt	flossing wrist	If flossing with pick, we mark the flossing wrist as left or right – otherwise mark it as both wrists
fl_tp	flossing type	Flossing type is marked as string or pick