DoseMate: A Real-world Evaluation of Machine Learning Classification of Pill Taking Using Wrist-worn Motion Sensors

Antoine Nzeyimana ANZEYIMANA@UMASS.EDU
University of Massachusetts, Amherst, USA

Anthony Campbell ANTHONYCAMPB@UMASS.EDU
University of Massachusetts, Amherst, USA

James M. Scanlan JAMES.SCANLAN@PROVIDENCE.ORG
Swedish Health Services, Seattle, USA

Joanne D. Stekler JSTEKLER@UW.EDU
University of Washington School of Medicine, Seattle, USA

Jenna L. Marquard MARQUARD@UMN.EDU
University of Minnesota, Minneapolis, USA

Barry G. Saver BARRY.SAVER@SWEDISH.ORG
Swedish Health Services, Seattle, USA

Jeremy Gummeson JGUMMESO@UMASS.EDU
University of Massachusetts, Amherst, USA

Abstract

Non-adherence to medication is a complex behavioral issue that costs hundreds of billions of dollars annually in the United States alone. Existing solutions to improve medication adherence are limited in their effectiveness and require significant user involvement. To address this, a minimally invasive mobile health system called DoseMate is proposed, which can provide quantifiable adherence data and imposes minimal user burden. To classify a motion time-series that defines pill-taking, we adopt transfer-learning and data augmentation based techniques that uses captured pill-taking gestures along with other open datasets that represent negative labels of other wrist motions. The paper also provides a design methodology that generalizes to other systems and describes a first-of-its-kind, in-the-wild, unobtrusively obtained dataset that contains unrestricted pill-related motion data from a diverse set of users.

Data and Code Availability The dataset and machine learning code from our primary study is released at https://gitlab.com/umass-smelt-lab/dosemate. The existing open datasets used in our study can be found at their respective repositories; FIC: https://zenodo.org/records/4421861; IM-WSHA: https://github.com/SheikhBadaruddinTahir/IM-WSHA; Cafeteria: http://cecas.clemson.edu/~ahoover/allday/; Handy: http://www.baskent.edu.tr/~hogul/handy; Daily Sports, MHEALTH and PAMAP2: https://archive.ics.uci.edu/datasets.

Institutional Review Board (IRB) The study that collected the data presented here was approved by the University of Massachusetts Amherst IRB (Study #1286) and Providence Health and Services (Study #SWD5984S-16).

1. Introduction

Many life-saving medications require longitudinal adherence to a prescribed regimen to achieve maximal efficacy. Yet, achieving high levels of adherence is often difficult for patients (Marcus et al., 2016). Suboptimal adherence to medications was estimated to cost $100 billion/year in the United States almost 2 decades ago (Osterberg and Blaschke, 2005) and a more recent estimate suggested annual costs in the neighborhood of $500 billion (Watanabe et al., 2018).

Improving medication adherence requires new methods of measuring and enhancing adherence that are both economically and behaviorally sustainable over the long term, since the benefits of many medications (e.g. medications for controlling blood pressure

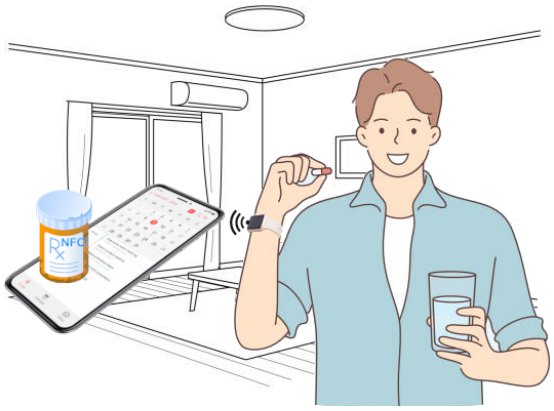

Figure 1: DoseMate is an unobtrusive medication adherence system used to quantitatively track pill intakes through classification of motion data collected from a wrist-worn device.

and diabetes, ART and PrEP for HIV) accrue with consistent levels of adherence that are sustained over time. Non-adherence is a complex behavioral issue with many drivers. The evidence for interventions that improve medication adherence is quite modest, at least in part due to the complexity of the drivers of non-adherence (Nieuwlaat et al., 2008; Cross et al., 2016). For cancer oral agents, a review found limited evidence that therapeutic educational interventions improve adherence (Arthurs et al., 2015).

A diverse set of methods for improving adherence have been proposed that range from simple pen-and-paper journaling routines to instrumented pill bottles that track opening and closing events, updated with the most recent time a pill was taken, to ingestible sensors powered by stomach acid-activated batteries, embedded in pills (Spinelli et al., 2020). While these efforts represent steps towards understanding and improving adherence for patients, they require specialized hardware that may not be cost-effective for large patient populations. Furthermore, these systems have no way to longitudinally and autonomously detect pill ingestions, which can lead to under-reporting. On the other hand, solutions that require heavy user-involvement have the risk of further reducing existing adherence levels. In summary, there remains a strong need for a light-weight solution that provides quantifiable adherence data that maximizes performance and user satisfaction while minimizing cost.

In this paper, we describe the design, development, deployment, and evaluation of a mobile health system called DoseMate, that offers the potential to improve both measurement and enhancement of adherence in a generalizable, sustainable, and scalable manner. The DoseMate system uses a combination of inexpensive, flexible, non-invasive technologies to trigger and measure pill-taking accurately, and in a proof-of-concept study, is intended to facilitate improved adherence to HIV-related medications. The DoseMate system provides timely reminders and real-time feedback to patients and their providers in a manner that is economically and behaviorally sustainable.

To demonstrate the potential of artificial intelligence for pill ingestion detection, we developed models trained on data collected using our system. We analyze a dataset collected with our system that contains over 1000 instances of positively labeled pill ingestion events from a clinical population of 20 users. This data was collected over the span of 6 months in-the-wild per participant, through an IRB-approved research study. The key challenge in using this data is the lack of clear labels that denote specific gestures within a motion time-series that denote actions that collectively define pill-taking, including bottle opening, bottle manipulation, pill manipulation, and hand-to-mouth gestures for pill ingestion and liquid consumption. Collecting labels for these individual gestures within the overal time-series would be too burdensome to collect in the wild from the vulnerable user population of our study. Furthermore, by design, our system captures only positively labeled events, which results in an imbalanced dataset that requires negative training examples from other potential activity classes to train machine learning (ML) models that detect pill-intake events. To overcome this obstacle, we adopt transfer-learning and data augmentation based techniques that uses other open datasets that represent other wrist motions that can be similar to pill intakes (i.e. eating) and other daily living activities that may be vastly different (i.e. walking, exercise). Our key contributions are:

- A design methodology for a minimally-invasive mobile smart calendar system that facilitates the collection of labeled motion data from medication intakes with minimal burden imposed on users. Users need only tap a pill bottle instrumented with an NFC tag to initiate data collection from a wrist-worn motion sensor and respond to a confirmation message with one tap.
- A first-of-its-kind, in-the-wild, unobtrusively obtained dataset that contains over 1000 unre-

stricted pill-related motion examples from a diverse set of users. The data is collected from either the dominant or non-dominant wrist of a user and contains a subset of different activities related to pill-taking including hand to mouth gestures and bottle and pill manipulation. This dataset differs from prior work in that it was collected in the naturalistic environment of the participant with no script or prompt from a study coordinator and represents pill taking motions present in a typical day-to-day routine.

- The design and evaluation of feature-based machine learning (ML) and deep learning models for pill intake recognition using wrist-worn motion sensor data from users in their free living environments. We achieve an average balanced accuracy 94% and 95% across different permutations of training datasets for feature-based ML and deep learning models respectively.

2. Rationale for the DoseMate System

DoseMate is a data collection system for acquiring wrist motion data from users to train machine learning models. DoseMate presents itself to the user as a smart calendar app that provides a minimally invasive medication logging system; the objective of the system is to maximize the likelihood that motion data related to intakes is accurately captured while remaining immediately useful to the user. Given the potential negative consequences in giving a patient incorrect adherence feedback, we have designed our system around lightweight reminders that should only improve adherence outcomes while collecting motion data in the background during pill-taking activities.

2.1. NFC-based logging

Pill bottles are instrumented with Near Field Communication (NFC) tag stickers that interact with a smartphone-based calendar system. The purpose of including these tags is to: (1) Ensure that the patient is taking the correct medication if their regimen includes multiple medications and (2) Reinforce behavior of patients to bring their medications with them to the location where they will ultimately be ingested. After reading the NFC tag, the patient is prompted to log their intake for the corresponding medication. If an incorrect or invalid medication is read, the system warns the patient. In our data collection system, this tag interaction is also used to initiate data capture.

2.2. Time-based logging

Similar to other medication journaling systems, such as the one recently implemented as part of the Apple HealthKit framework[1], DoseMate provides time-based notifications based on the ideal-time for taking a particular medication. These time-based reminders are implemented using Android's alarm API and deliver reminders as notifications on a mobile phone that are mirrored on a wearable. The notifications prompt the user to take the medication and includes a button for manually logging the intake if the patient does not have access to the paired NFC tags.

2.3. Background Data Collection

After being triggered by an NFC tap gesture, DoseMate collects 6-axis motion data (i.e. Gyroscope, Accelerometer) from a smartwatch. This data is meant to reflect wrist motions related to pill taking activities including bottle opening, pill transfer from bottle-to-hand, hand-to-mouth motions, fluid intake, and bottle closing. After collection, the data is transferred to a backend server for subsequent analysis. Our ultimate vision is to use models trained on this data to continuously monitor user motions in the background that autonomously reminds a user to log an intake with no NFC tap used to initiate the process.

3. Related Work

In this section we provide an overview of related techniques used to track pill ingestion events. This research focuses on data science as well as software and hardware systems developed in the ubiquitous computing and mobile computing research communities.

3.1. Tracking pill ingestion events

The ubiquituous computing and mobile systems research community has explored tracking pill ingestion using technology-based solutions. One class of these solutions focuses on augmenting pills themselves to understand whether or not a pill has been ingested (Chai et al., 2022b). A recent study focused on

1. https://support.apple.com/en-us/105064

tracking pill ingestions of PrEP patients using RFID-enhanced pills (Chai et al., 2022a). While smart pills offer definitive proof of medication adherence, they do not scale cost-effectively and would be difficult to sustain for longitudinal adherence. Furthermore, additional hardware results in additional barriers such as the reader used for RFID-enhanced pills (Chai et al., 2022a). In our system, we record motion data surrounding a pill-intake to provide evidence that a pill was likely taken using a commercial smartwatch that can be used for all normal smartwatch functions, with predictions computed by a machine learning classifier. More closely related to our work are systems that use combinations of motion features, cameras, and machine learning to estimate pill ingestion (Chen et al., 2014; Moccia et al., 2022; Wang et al., 2014; Cheon et al., 2020; Huynh et al., 2009). While these types of systems are more scalable than smart pill solutions, we note that previous datasets have been limited to lab-based data collection efforts that do not provide a good sense of performance from unrestricted and unprompted pill taking behavior in-the-wild. Instead we opted for a non-invasive (i.e. no cameras) and minimally disruptive system that uses a smartphone and smart watch to collect data from users seamlessly during their day to day activities.

3.2. Activity recognition using features extracted from motion data

A growing body of work has looked at measuring hand-to-mouth gestures related to activities with high impact including eating (Dong et al., 2013), drinking (Hamatani et al., 2018), brushing (Akther et al., 2021), smoking (Parate et al., 2014), and other "bad habits" (Shoaib et al., 2015). These efforts have been able to exploit periodicity of hand-to-mouth motions to increase classification performance, where one-off gestures can be easily filtered since smoking, drinking, or eating involve a series of similar motions during a session that defines the activity class. Tracking pill-intakes is a substantially different problem in that a pill-intake is typically a one-shot action. Instead of looking at repetitions of motion, we look at the structure of pill-related gestures that contain several activities that include bottle opening, pill-manipulation, and hand-to-mouth motions for pill and fluid intake. We capture diverse data across users to train generalized classifiers that can identify permutations and variations of pill-related motions.

4. DoseMate: An Unobtrusive Pill Intake Tracking Platform

DoseMate is a data-driven, mobile smart calendar system that was developed for phones using Android OS $\geq$ 9.x and smartwatches running Android WearOS 2.x. An overview of the system is depicted in Figure 2. The system was developed, tested, and deployed over a 5 year period from 2017 - 2022 and was used to successfully enroll a cohort of 20 users for an in-the-wild, IRB-approved, clinical study that collected over 1000 examples of positively-labeled medication intake gestures from unconstrained users during their normal routines. In this section, we describe each system component and how they collectively contribute towards the aim of quantifying medication adherence using machine learning techniques that correctly classify wrist motions as pill intakes.

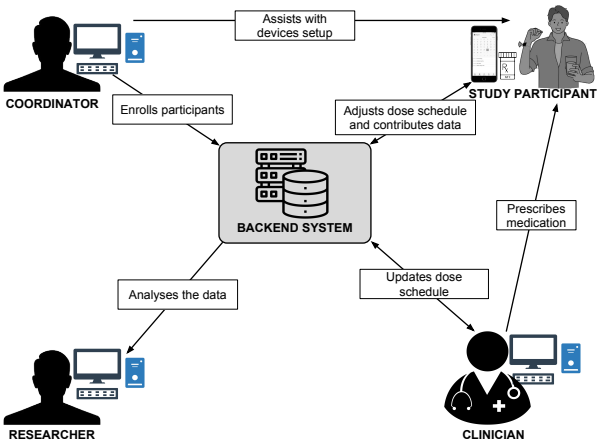

Figure 2: DoseMate system overview

4.1. System Life Cycle

The hardware and software used in our system is designed, and was co-developed with clinicians, to be relatively simple for a clinician and research coordinator to configure. We use commercial off-the-shelf (COTS) hardware components wherever possible with a custom Android 9.x app and a backend implemented using the Flask web micro framework[2]. The system was used to successfully complete a clinical study during the COVID-19 pandemic in North America with researchers and clinicians on opposite coasts. We describe this process as a general medication adherence system and later sections describe study-specific details.

2. https://flask.palletsprojects.com/

Two different types of data are collected with DoseMate – the first is data from user interactions with the smart calendar, while the second is mostly invisible to the user and is motion data that quantitatively describes individual pill intakes. We describe these in further detail. After enrollment, a participant interacts with DoseMate. The visible portion of the app is a smart calendar with time and interaction based notifications (Figure 3), with motion data seamlessly collected in the background.

Interactions with Smart Calendar The home screen of the application is a smart calendar, graphically depicting adherence during the current calendar month where green indicates doses all in the on-time window, yellow indicates some doses outside of the on-time window or missing doses, and orange (colors selected to avoid issues with color blindness) indicates all doses are missing. We note this synthetic user would be one struggling with adherence. Study participants get a notification on the phone and watch as depicted in Figures 3 (b). Dose status, along with potential updates to the dosing schedule, are synchronized with the backend server using a restful API.

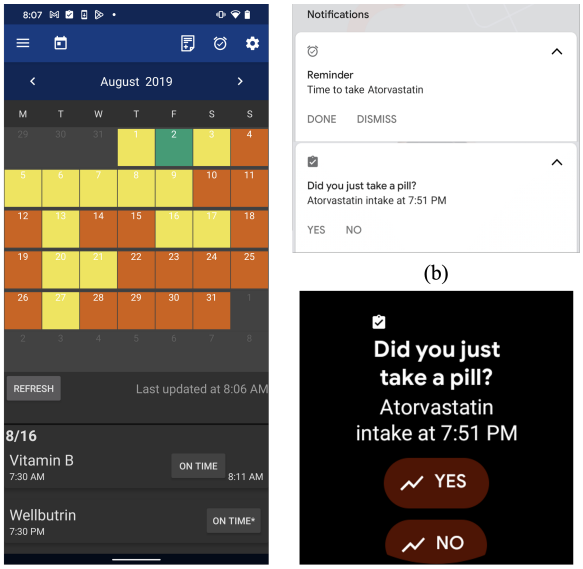

Figure 3: DoseMate is implemented as a smart calendar app (a). Study participants interact with the app through the calendar screen and in response to notifications on the phone (b) or watch (c).

Wrist-Motion Data Collection Wrist motion data is collected after tapping a pill bottle with an NFC tag against the mobile phone – if a dose for this medication already exists or is outside the early/on-time/late window, the participant is warned that this is an extra dose. If inside a valid window with no recorded dose, this event triggers data collection on the watch – the watch generates haptic feedback to indicate it has started recording and 19 seconds later, the user receives a confirmation depicted in Figures 3 (b) and (c). The assumption is that the user will have finished pill taking activities during this window; the window size of 19 seconds was empirically determined during a round of unprompted lab-based data collection. These activities can include bottle manipulation (movement, opening), pill manipulation, hand-to-mouth gestures (pills and/or fluid intake), and possibly others. After this time interval elapses, the user confirms that they did indeed take a pill or that they did not take a pill. The latter rarely occurs but is particularly useful when users are first learning to use the system to avoid mis-labeled data. In practice, data labeled as not-taken was used as negative class training data; this data was provided by several volunteer participants and data during the training period was discarded.

5. Using Machine Learning to Recognize Pill Ingestion

Accurately tracking and confirming pill intake events can be beneficial for users who are at risk of missing or over-using their medication. While there exist other tools that are useful in supporting medication adherence, such as calendar applications and timer-equipped pill bottles, using smartwatch sensors provides additional benefits. First, leveraging wrist-mounted motion sensors can provide a physical confirmation signal that the medication has been taken. Second, that signal can be aligned with medication intake schedules on smart calendar applications to limit the number of false positives.

However, accurately detecting pill intake events using motion sensors presents a number of challenges. First, the data collected from inertial measurement units (IMU) in free living conditions is usually noisy, and it is difficult to reconstruct the actual hand trajectory. Second, there are inter-user variations in medication intake dynamics which make it challenging to find consistent common patterns across all users. Third, there are other hand to mouth gestures, such as eating and drinking, that are similar to pill intake gestures which can cause false positive

classifications. In addition to these challenges, the dataset collected from real users taking medication is typically small and class-imbalanced because the vast majority of user-provided labels are for positive intakes within the context of the smart calendar app.

In order to cope with the above challenges, we explore the use of two machine learning approaches to model user gestures while taking medication. We apply both feature engineering-based machine learning and more recent deep learning techniques. The task is modeled as a time-series classification problem. The input is a multivariate time-series of accelerometer and gyroscope readings over a $\sim$ 19 second time interval and the prediction is a binary label indicating whether the user took the medication or not.

5.1. Dataset Description

Our main dataset (DoseMate) comes from 20 participants enrolled in an IRB approved study in the United States. The data was collected from the medication-taking participants between January 2020 and June 2022. The main dataset we use includes 3-axis accelerometer and gyroscope data from Android watches worn by the participants. During the scheduled medication intake time interval, the participants tapped the NFC-equipped pill bottle to a smartphone to indicate a possible beginning of a medication intake event. Then, the smartphone communicates with the smartwatch worn by the participant to start recording motion data. After 19 seconds, the recording stops and the participant is prompted to provide a label indicating whether the medication was taken or not. The data from all participants was anonymized and aggregated in a backend database. A total of 1154 medication intake candidate events were recorded including 185 labeled as non-intakes (negative labels), each lasting $\sim$ 19 seconds.

In order to properly evaluate the effectiveness of medication intake recognition algorithms, we benchmark and train our models with existing open datasets from wrist-worn IMU sensors in addition to our primary dataset. This is useful because our primary dataset is very class-imbalanced. While the open datasets were collected from different environments using different sensor settings, they still make a reasonable benchmark because they include a diverse range of physical activities that one would want to differentiate from pill intakes for a practical recognition system. Also, comparing against these different datasets provides an insight into which type of activities are easier and which ones are harder to differentiate from pill intakes. We ensure all datasets, including our primary dataset, are re-sampled at the same sampling rate of 15 hertz, segmented into 19-seconds segments, and go through the same pre-processing and feature extraction steps. This helps to alleviate any bias that would come from sensor specificities rather than the actual motions we want to model. Table 1 provides a summary of the datasets we used.

5.2. Insights from raw IMU data

To provide insight into the datasets we captured from users in the wild, we collected a small subset of data in the lab with video recordings used as ground truth and labelled different gestures within the intake window. The purpose of this data collection was to provide insights and inspiration into what features could be potentially useful for feature-based machine learning. We illustrate this data in Figure 4, showing the most prominent axes of motion. Here, we show the start and end times of events observed during the data collection period including bottle openings and closings, pill to mouth motion, and drinking.

5.3. Feature extraction and feature-based classification

The input data to our classification pipeline are the approximately 19 second segments comprising 3-axis accelerometer time-series and 3-axis gyroscope time-series resampled at 15 Hz. The output is a binary label indicating whether the segment contains a pill intake event or not. Our feature extraction pipeline is depicted in Figure 5.

Before computing the features used for feature-based classification, we augment the time-series with a filtered version of each of the accelerometer time-series. We apply both a low-pass filter with a cut-off frequency of 2 Hz and a high-pass filter with a cutoff frequency of 0.2 Hz. After adding the filtered time-series, we also compute and add the absolute value of each time-series axis and the root-mean-square of the accelerometer and gyroscope 3 axes time-series.

After time-series augmentation, we proceed with feature computation comprising global features of the whole segment and statistics of 4 sub-segments of the whole 19 seconds. Global features include real components 1 to 5 of the fast Fourier transform (FFT) of each time-series and the detrended fluctuation analysis (DFA) (Bryce and Sprague, 2012) exponent α of the raw IMU time-series. These global features help

Dataset	# Train	# Dev	# Test	Sampling rate (Hz)	Min/Max/Avg Len (sec)
DoseMate* (This work)	924 / 149	115 / 18	115 / 18	26	18.9 / 19.5 / 19.2
Cafeteria Shen et al. (2016)	3605	451	451	15	14 / 30 / 19.2
Daily sports (left arm) Altun et al. (2010)	1824	228	228	25	19 / 19 / 19
Daily sports (right arm) Altun et al. (2010)	1824	228	228	25	19 / 19 / 19
FIC Kyritsis et al. (2019)	679	85	85	100	19 / 19 / 19
Handy Açıcı et al. (2018)	691	86	86	52	19 / 19 / 19
MHealth Banos et al. (2014)	918	115	115	50	19 / 19 / 19
Pamap2 Reiss and Stricker (2012)	1078	135	135	100	19 / 19 / 19
Wsha Tahir et al. (2020)	173	22	22	24	19 / 19 / 19

Table 1: Number of examples across different datasets used for evaluating our machine learning models. * The data from DoseMate shows the total number of examples along with the number of negative examples in each set. Examples from the open datasets are all considered as negative (non-intakes).

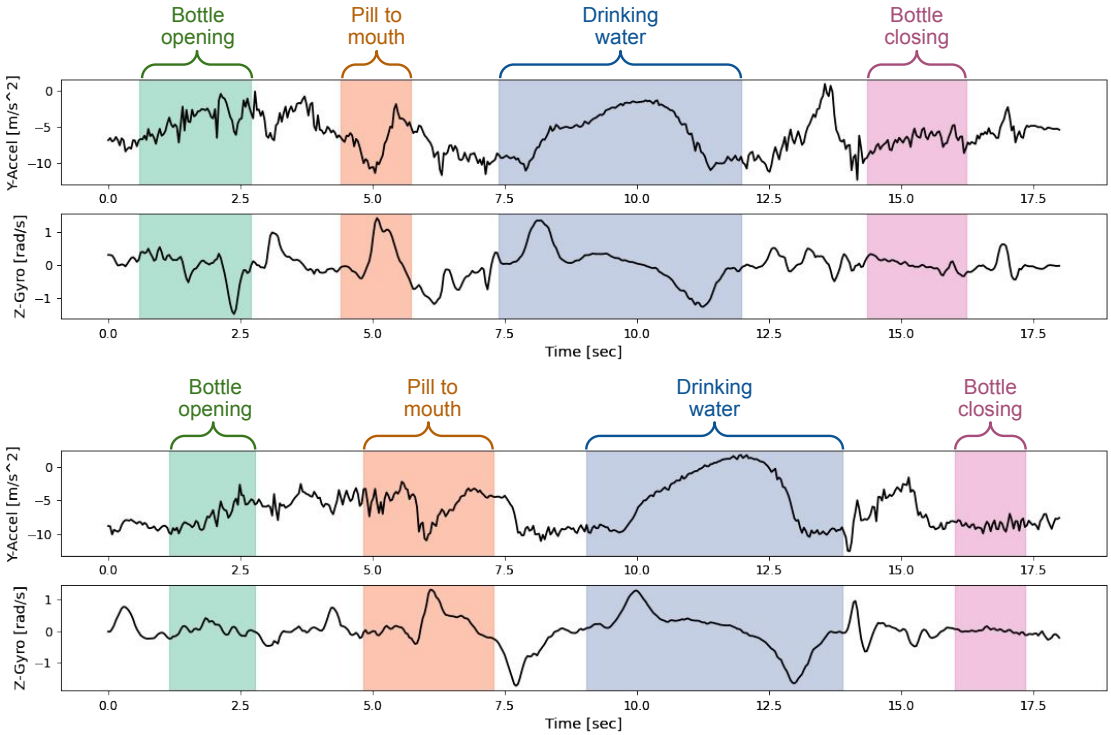

Figure 4: Two examples of annotated pill intake time-series from two prominent axes of the IMU

capture some of the periodicity and level of fluctuations present in the time-series. Furthermore, we split the time-series into four sub-segments of equal lengths and compute statistics of each sub-segment as features. These splits are motivated by the observation that most pill intakes included at least four distinctive intervals corresponding to pill bottle opening, pill ingestion with a hand to mouth gesture, liquid drinking and a final motion which might be a pill bottle closing or any other concluding motion. For each of the time-series in each sub-segment, we computed the mean, standard deviation, median, mean-median difference and the interquartile range (IQR) as sub-segment features.

In total, 710 features are computed and we perform feature ranking based on mutual information (Cang and Partridge, 2004). We evaluate feature-based ensemble classifiers including Random

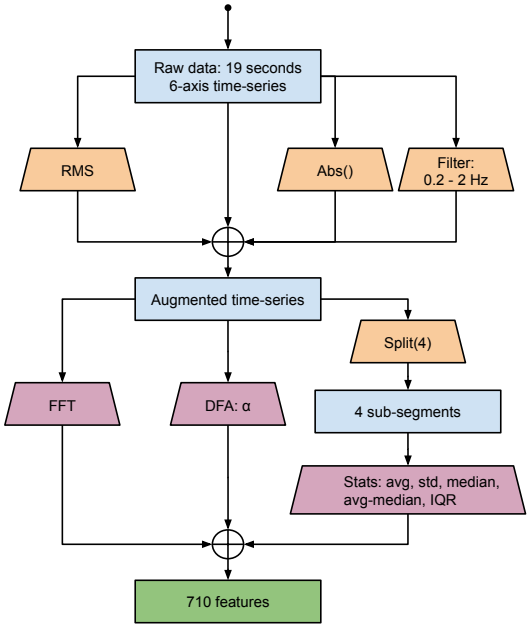

Figure 5: Feature extraction pipeline.

Forests (Breiman, 2001), Gradient Boosted trees with XGBoost (Chen and Guestrin, 2016) and balanced Bootstrap aggregating (Bagging) (Hido et al., 2009). In addition to the 710 features, we evaluate each classifier using the top ranked 100, 50, 25, 10 and 5 features as a model selection mechanism.

5.4. Using deep learning techniques

We explore the use of deep learning models for time-series classification as a way to detect pill intakes. Deep learning techniques are attractive because they alleviate the need for feature engineering when processing unstructured data such as images and sequences. Specifically, we use architectures based on recurrent neural networks (RNN), one-dimensional convolutional neural networks (CNN) and the transformer architecture (Vaswani et al., 2017).

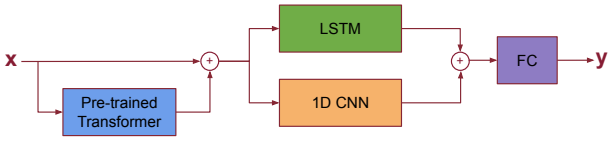

Figure 6: Deep learning-based architecture for IMU time-series classification. $\mathbf{x}$ is the multi-variate time-series input data and y is the classification label.

We use an Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) – fully convolutional network (LSTM-FCN) (Karim et al., 2017) as the base architecture for our exploration. The multivariate time-series data are processed by both the LSTM and the one-dimensional CNN (1D-CNN) in parallel. The outputs of the LSTM and the CNN are then concatenated and passed to a fully-connected module and a softmax function for final classification. We also augment the LSTM-FCN network with a Transformer (Vaswani et al., 2017) encoder module to allow for transfer learning through self-supervised learning (SSL) (Ericsson et al., 2022). Specifically, we follow the denoising task proposed in Zerveas et al. (2021) to pre-train our transformer module. The denoising task consists of randomly corrupting the input data by replacing small sub-segments of the time-series with zero values, then using the transformer to learn a function that predicts the original data for the corrupted sub-segments. Overall, we explore three architectural choices for the deep learning models:

(a) We process the input time-series x only using an LSTM-FCN network.

(b) We use a pre-trained transformer module to encode the input x before passing it to the LSTM-FCN network.

(c) We encode the input with the pre-trained transformer module and then concatenate the transformer output with the original input as a data augmentation mechanism before passing the resulting time-series to the LSTM-FCN network.

6. Evaluation

6.1. Experimental setup

We evaluate our feature-based classifiers and deep learning models using two types dataset setups:

1. **DoseMate only**: we only use data from study participants, having few negative examples.

2. **DoseMate + 1 external dataset**: we combine DoseMate data with each external dataset as negative examples. Each external dataset is used to evaluate how trained classifiers can differentiate between pill taking events from our primary study and other hand motions recorded in the external dataset. The external dataset

is not used for augmenting our primary dataset, but for evaluation against it.

In each case we report the machine learning performance on the test set using specifity, balanced accuracy (BA) which is the arithmetic mean of sensitivity and specificity, and area under the ROC curve (AUC). Due to the imbalanced class labels, our main objective is to achieve high balanced accuracy.

We ran the machine learning experiments using Python 3.9. The feature-based machine learning pipeline uses scikit-learn and XGBoost packages and their dependencies. Deep learning experiments are implemented using PyTorch version 1.3. We randomly split the data into training, development and test sets in ratios of 80%, 10% and 10% respectively, and ensured the same ratio of negative examples in each set. For feature-based models, model hyper-parameters were tuned on a development set split with a coarse-grained grid search. For deep learning, we used common hyper-parameter settings found in literature. These hyper-parameter settings are presented in B. We ran each of our experiments for 10 runs and we report the average measures. Our experiments were run on a university computing cluster using Nvidia RTX 2080 Ti graphic processing units (GPU). It takes four minutes to run one round of XGBoost training on all data setups using one GPU, while it takes seven minutes for the LSTM-FCN network on one GPU. Pre-training the Transformer encoder on the de-noising task for a total of 200 epochs takes four hours. We use the same noise parameters proposed in (Zerveas et al., 2021). For the larger model of Transformer+LSTM-FCN with input concatenation, it takes approximately one hour to train (i.e. fine-tuning) on all setups (20 epochs per setup).

6.2. Machine learning performance

The best results for the feature-based machine learning are presented in Table 2 where we use the top ranked 50 features and XGBoost classifier. All open datasets combined with DoseMate dataset result in higher performance than using DoseMate dataset only. Using time-series from the Cafeteria dataset as negative examples results in the highest performance, achieving up to 98% balanced accuracy. This Cafeteria dataset may also be easier to distinguish from pill intakes as participants in that study (Shen et al., 2016) make many short food intake gestures followed by longer resting intervals. Considering all datasets, we achieve 94% balanced accuracy on average.

The best results of the deep learning models are presented in Table 2. We observe that relative variations across all data setups are similar to the feature-base machine learning results. This indicates that there are inherent patterns in the datasets that make it possible to distinguish pill intake events from non-intake examples. However, using the DoseMate only dataset results in worse performance than XGBoost. Overall, the deep learning model achieves 95% balanced accuracy, which is comparable to XGBoost.

We present ablation results in Table 3 and Table 4. There are four options for the Transformer+LSTM-FCN model: whether to concatenate the transformer output with the input or not, and whether to fine-tune the transformer weights or not. Using transfer learning with fine-tuning achieves better results for deep learning models. On average, using deep learning with transfer learning achieves better results than feature-based classifiers.

6.3. Feature importance analysis

With feature-based machine learning, we are able to analyze the importance of different features to the classification task. In our experiments, the XGBoost library provides a feature importance assessment based on how many times each feature gets split on by the decision tree estimators. Based on the nine evaluation setups presented in Table 2, we averaged the importance scores computed by XGBoost for each feature in each setup and then ranked them. We present the top 10 features in Figure 7.

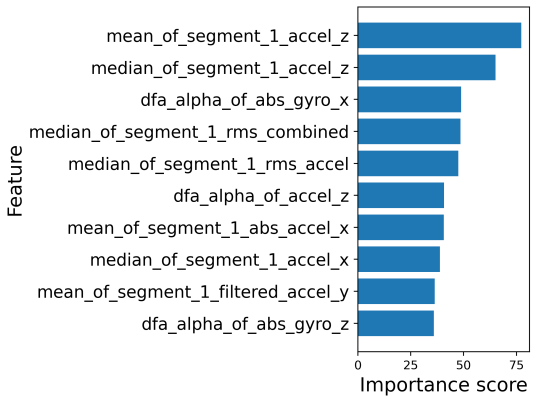

Figure 7: Top 10 features ranked used by XGBoost. The feature importance score indicates the average number of decision tree splits on each feature.

Table 2: Feature-based machine learning versus deep learning results. For the feature-based setup, we report results from XGBoost model and top ranked 50 features. The deep learning model uses the architecture in Figure 6 where the input is passed to the pre-trained Transformer without input concatenation and no fine-tuning of the Transformer. (SP: Specificity, BA: Balanced accuracy, AUC: Area under the ROC curve). The DoseMate dataset includes only data from study participants. Instances of open datasets are used as negative examples (no intakes) in addition to those in the DoseMate dataset. Bold results indicate the best metric between XGBoost and the deep learning model for a given setup(row-wise comparison for each metric)

	XGBoost			Deep learning		
Dataset setup	**SP**	**BA**	**AUC**	**SP**	**BA**	**AUC**
DoseMate only	.83	**.90**	**.93**	.83	.88	.92
DoseMate+ Cafeteria Shen et al. (2016)	.98	.98	1.0	**.99**	.98	1.0
DoseMate+ Daily sports (left arm) Altun et al. (2010)	**.96**	**.96**	.99	.95	.95	.99
DoseMate+ Daily sports (right arm) Altun et al. (2010)	.95	.96	.99	**.98**	.96	.99
DoseMate+ FIC Kyritsis et al. (2019)	.90	.95	.98	**.97**	**.96**	.98
DoseMate+ Handy Açıcı et al. (2018)	.91	.94	.98	**.96**	**.95**	.98
DoseMate+ MHealth Banos et al. (2014)	.90	.94	.99	**.97**	**.97**	.99
DoseMate+ Pamap2 Reiss and Stricker (2012)	.91	.94	.99	**.98**	**.97**	.99
DoseMate+ IM-WSHA Tahir et al. (2020)	.90	**.93**	.96	**.92**	.92	.96
Average	.92	.94	.98	**.95**	**.95**	.98

Table 3: Comparison of average scores of different feature-based classifiers. SP: Specificity, BA: Balanced accuracy, AUC: Area under the ROC curve

Model	**SP**	**BA**	**AUC**
XGBoost	**.92**	**.94**	**.98**
Random Forest (RF)	.88	.93	.97
Balanced RF	.89	.93	.97
Balanced Bagging	.90	**.94**	.97

Table 4: Comparing average scores of different deep learning model setups. We have options to concatenate (CC) the transformer output with the input or not, and to fine-tune (FT) the transformer or not. The best results in bold. SP: Specificity, BA: Balanced accuracy, AUC: Area under the ROC curve

Model	**FT**	**CC**	**SP**	**BA**	**AUC**
Trfm. + LSTM-FCN	Yes	No	**.95**	**.95**	**.98**
Trfm. + LSTM-FCN	Yes	Yes	**.95**	**.95**	**.98**
Trfm. + LSTM-FCN	No	Yes	.94	.94	**.98**
Trfm. + LSTM-FCN	No	No	.93	.92	**.98**
LSTM-FCN	n/a	n/a	.94	.94	**.98**

By observing the top 10 features presented in Figure 7, we make the following observations:

1. 7 out of the 10 top features relate to statistics on the first quarter/segment of the time-series. This generally may correspond to pill bottle opening and pill-to-mouth gestures. Therefore, it is not surprising that this segment contains the most distinctive features as opposed to other arbitrary gestures observable from the wrist-worn IMU.

2. The top two features are about the average of Z-axis accelerations. This might correspond to the lifting gesture, which may also correspond to the hand to mouth gesture.

3. Some of the top features also correspond to the RMS signal of the first quarter/segment which indicates distinctive energy of the time-series during that interval.

4. There are also features corresponding to the detrended fluctuation analysis (DFA) exponent (α), which indicates distinctive statistical self-affinity of the signals. This feature typically measures the roughness (fluctuations) or smoothness of the time-series.

7. Discussion

This manuscript has described a first-of-its-kind system that unobtrusively collects pill-taking gestures from study participants in the wild. Our system represents a first step towards low cost, scalable, quantitative adherence measurement, but there are many open challenges to be explored in this domain. We highlight some of these in this section.

Explainability. While the use of deep learning for activity recognition does not lend itself to easy explanation from the human, it is based on known principles that work well in practice such as universal function approximation and regularization. This makes it possible for models to generalize to unseen data samples in practice, as shown in 6.2. However, the lack of easy explainability is not limited to neural networks. Even feature-based machine learning algorithms such as random forest are not fully explainable. While we are able to rank various features used and intuitively confirm their effectiveness, the machinery of combining many features in a random forest model with hundreds of estimators is not also fully explainable.

Previous studies involving wrist-worn motion sensors such as (Parate et al., 2014) attempt to reconstruct motion trajectories. This can allow easy visualization of the 3-dimensional motion, enable the use of displacement and rotation features and convey some level of explainability. However, this is most practical for simple and constrained motions such as short hand-to-mouth gestures (e.g. smoking detection in Parate et al. (2014)) and often requires 9-axis IMUs with motion processing units that output motion data as quarternions. The aggregate tracking error for a 6-axis IMUs on commodity smartwatches and the fact that pill intake events span more than 12 seconds involving multiple and diverse gestures make it less practical in our case (Shen et al., 2018).

Online Classifier Performance: In this work, we performed a post-facto analysis of classifiers trained on data collected during a study. A future implementation of an adherence system would deploy the classifier on a mobile phone or smartwatch and continuously record motion data in the background. Rather than requiring an NFC tap or other prompt to initiate data collection, this approach would classify any motion segment as an intake or non-intake and prompt the user to record it. Running the classifier continuously will potentially trigger false-positive intake logging prompts. It remains unknown whether these would be infrequent enough to remain unbothersome to users while avoiding more dangerous false-negatives. Ideally, a classifier could be made accurate enough to autonomously identify intakes with no user intervention while contextual optimizations can limit when classifications occur (i.e. during valid scheduled intake windws); however, significant work remains to achieve the needed accuracy and confidence in the classifiers to realize this vision. Furthermore, additional steps would be needed to optimize the power attributed to data capture and inference.

Limitations of using external datasets: While we relied on external IMU datasets to alleviate the problem of class imbalance in our primary, using external datasets collected from various sensors and environments brings different problems. First, there is a potential distributional shift caused by sensor heterogeneity since various sensors have different amount and types of noise associated with them. Second, sensor heterogeneity means that we have to resample the time-series to the lowest sampling rate among the datasets, causing further interpolation noise. Finally, while the datasets we used were all collected from wrist-mounted IMUs, there were also a large diversity in terms of activities performed by the users while wearing the IMUs. Therefore, all the above factors impact the performance observed when we consider these external datasets as negative examples. A viable approach to mitigate this limitation in future studies would be to only consider datasets from the same type of IMU as the one used in the primary study. One way to do this would be to either collect additional data in the background from the same primary study subjects or consider using a more widely used common wearable such as Android or Apple watches for which open IMU datasets exist.

8. Conclusion

In this paper we described the design, implementation, and evaluation of a mobile health system used to quantitatively assess and encourage medication adherence through accurate classification of pill intakes. To address the lack of clear labels that denote specific gestures within a motion time-series that define pill-taking, we adopted transfer-learning and data augmentation based techniques that used other open datasets that represented negative labels of other wrist motions. Our results show feature-based machine learning achieved an average F1 score of 96.8%, while deep-learning models achieved an average F1 score of 96.2% across different training datasets that contain different combinations of similar and disparate datasets.

Acknowledgments

This research was funded by NIMH award numbers R01MH109319 and 5R21MH126326-02.

References

Koray Açıcı, Çağatay Berke Erdaş, Tunç Aşuroğlu, and Hasan Oğul. Handy: A benchmark dataset for context-awareness via wrist-worn motion sensors. *Data*, 3(3):24, 2018.

Sayma Akther, Nazir Saleheen, Mithun Saha, Vivek Shetty, and Santosh Kumar. Mteeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(2), jun 2021. doi: 10.1145/3463494. URL https://doi.org/10.1145/3463494.

Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

Gilly Arthurs, Janice Simpson, Andrea Brown, Ohnma Kyaw, Sharon Shyrier, and Catherine M Concert. The effectiveness of therapeutic patient education on adherence to oral anti-cancer medicines in adult cancer patients in ambulatory care settings: a systematic review. *JBI Evidence Synthesis*, 13(5):244–292, 2015.

Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6*, pages 91–98. Springer, 2014.

Leo Breiman. Random forests. *Machine learning*, 45: 5–32, 2001.

RM Bryce and KB Sprague. Revisiting detrended fluctuation analysis. *Scientific reports*, 2(1):315, 2012.

Shuang Cang and Derek Partridge. Feature ranking and best feature subset using mutual information. *Neural Computing & Applications*, 13:175–184, 2004.

Peter R Chai, Yassir Mohamed, Maria J Bustamante, Georgia R Goodman, Jesse Najarro, Jose Castillo-Mancilla, Alejandro Baez, Olivia Bronzi, Matthew C Sullivan, Luis M Pereira, et al. Digiprep: a pilot trial to evaluate the feasibility, acceptability, and accuracy of a digital pill system to measure prep adherence in men who have sex with men who use substances. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 89 (2):e5, 2022a.

Peter R Chai, Clint Vaz, Georgia R Goodman, Hannah Albrechta, Henwei Huang, Rochelle K Rosen, Edward W Boyer, Kenneth H Mayer, and Conall O'Cleirigh. Ingestible electronic sensors to measure instantaneous medication adherence: A narrative review. *Digital Health*, 8:20552076221083119, 2022b.

Chen Chen, Nasser Kehtarnavaz, and Roozbeh Jafari. A medication adherence monitoring system for pill bottles based on a wearable inertial sensor. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4983–4986. IEEE, 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Andy Cheon, Stephanie Yeoju Jung, Collin Prather, Matthew Sarmiento, Kevin Wong, and Diane Myung-kyung Woodbridge. A machine learning approach to detecting low medication state with wearable technologies. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4252–4255. IEEE, 2020.

Amanda J Cross, Rohan A. Elliott, Kate Petrie, Lisha Kuruvilla, and Johnson George. Interventions for improving medication-taking ability and adherence in older adults prescribed multiple medications. *The Cochrane database of systematic reviews*, 5:CD012419, 2016. URL https://api.semanticscholar.org/CorpusID:218555810.

Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover. Detecting periods of eating during free-living by tracking wrist motion. *IEEE journal of biomedical and health informatics*, 18(4): 1253–1260, 2013.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and chal-

lenges. *IEEE Signal Processing Magazine*, 39(3): 42–62, 2022.

Takashi Hamatani, Moustafa Elhamshary, Akira Uchiyama, and Teruo Higashino. Fluidmeter: Gauging the human daily fluid intake using smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3), sep 2018. doi: 10.1145/3264923. URL https://doi.org/10.1145/3264923.

Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5-6):412–426, 2009.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

H Hung Huynh, Jean Meunier, Jean Sequeira, and Marc Daniel. Real time detection, tracking and recognition of medication intake. *International Journal of Computer and Information Engineering*, 3(12):2801–2808, 2009.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data. *IEEE journal of biomedical and health informatics*, 2019.

Julia L Marcus, Leo B Hurley, Charles Bradley Hare, Dong Phuong Nguyen, Tony Phengrasamy, Michael J Silverberg, Juliet E Stoltey, and Jonathan E Volk. Preexposure prophylaxis for hiv prevention in a large integrated health care system: adherence, renal safety, and discontinuation. *Journal of acquired immune deficiency syndromes (1999)*, 73(5):540, 2016.

Sara Moccia, Sarah Solbiati, Mahshad Khornegah, Federica FS Bossi, and Enrico G Caiani. Automated classification of hand gestures using a wristband and machine learning for possible application in pill intake monitoring. *Computer Methods and Programs in Biomedicine*, 219:106753, 2022.

Robby Nieuwlaat, Nancy L. Wilczynski, Tamara Navarro, Nicholas Hobson, Rebecca A. Jeffery, Arun Keepanasseril, Thomas Agoritsas, Niraj Mistry, Alfonso Iorio, Susan M. Jack, Bhairavi Sivaramalingam, Emma C. Iserman, Reem A. Mustafa, Dawn Jedraszewski, Chris Cotoi, and R. Brian Haynes. Interventions for enhancing medication adherence. *The Cochrane database of systematic reviews*, 11:CD000011, 2008. URL https://api.semanticscholar.org/CorpusID:34782866.

Lars Osterberg and Terrence Blaschke. Adherence to medication. *New England journal of medicine*, 353 (5):487–497, 2005.

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

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. Closing the gaps in inertial motion tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 429–444, 2018.

Yiru Shen, James Salley, Eric Muth, and Adam Hoover. Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables. *IEEE journal of biomedical and health informatics*, 21(3):599–606, 2016.

Muhammad Shoaib, Stephan Bosch, Hans Scholten, Paul JM Havinga, and Ozlem Durmaz Incel. Towards detection of bad habits by fusing smartphone and smartwatch sensors. In *2015 IEEE international conference on pervasive computing and communication workshops (PerCom Workshops)*, pages 591–596. IEEE, 2015.

Matthew A Spinelli, Jessica E Haberer, Peter R Chai, Jose Castillo-Mancilla, Peter L Anderson, and Monica Gandhi. Approaches to objectively measure antiretroviral medication adherence and

drive adherence interventions. *Current HIV/AIDS Reports*, 17:301–314, 2020.

Sheikh Badar Ud Din Tahir, Ahmad Jalal, and Kibum Kim. Wearable inertial sensors for daily activity analysis based on adam optimization and the maximum entropy markov model. *Entropy*, 22 (5):579, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan H Watanabe, Terry McInnis, and Jan D Hirsch. Cost of prescription drug–related morbidity and mortality. *Annals of Pharmacotherapy*, 52 (9):829–837, 2018.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021.

Appendix A. System pre-deployment

Here we describe the steps needed to deploy an instance of DoseMate. These steps are outlined from the point of view of a medical clinic or hospital and assumes the backend system has been already installed by the researchers using appropriate computing, networking, and storage resources.

1. **Backend account creation:** The research coordinator creates study-specific and anonymous Google accounts for participants. Research coordinator creates a user in the DoseMate backend that links an anonymous ID with the Google account. The DoseMate app developer adds account information to the Google Play Store allowlist, configured in closed testing.
2. **Mobile software configuration:** The clinician enrolls patient in backend system using an anonymous identifier and the Google account. clinician enters the medication dose schedule which has flexible configuration according to time-of-day, day-of-week, and is assigned on a per-medication basis.
3. **Mobile device setup:** The research coordinator configures the software using a study-specific phone and downloads the smart calendar app from the Google Play store. Next, they pair the watch with the phone and install the companion app on the watch that runs a background service used for motion data collection.
4. **Mobile software configuration:** The research coordinator uses Google sign-in to securely connect the app with the back end system, which synchronizes the current medication schedule. Finally, they associate NFC stickers with each medication in the participant's dose schedule by tapping the phone against each tag and selecting an unassigned medication from a list of available medications. No information is stored in the tag memory and the tag's unique identifier is used to form an association with the medication.
5. **Participant enrollment:** The research coordinator meets with the participant and gives a walkthrough of the system components. The participant completes a survey that includes demographic information and details regarding current challenges with medication adherence.

The time series data in Figure 4 can see that there are some commonalities between the two intakes. During the pill intake, we observed analogous troughs in the gyroscope data and comparable increases in gyroscopic measurements along the z-axis. Likewise, the drinking action exhibited similar patterns. During the drinking period, the data collected from the accelerometer y-axis, we identified two parabolic formations and the gyroscope data for the z-axis displayed a consistently descending trajectory.

Appendix B. Deep learning model details and hyper-parameter setup.

Here we provide more details regarding the deep learning model architecture we used and the vari-

ous hyper-parameter settings we chose. As shown in Figure 6, the input time-series are first processed by the transformer module, which is feed-forward architecture based on the multi-headed self attention mechanism with residual connections across layers. The time-series are then processed in parallel by an LSTM module and a 1-dimensional convolution module. The outputs of the LSTM and convolutional module are then concatenated forming a hidden vector representation that is then passed to a fully-connected layer for final classification.

The hyper-parameters chosen for the deep learning model are presented in Table 5. The architectural dimensions were chosen based on common setups found in practice to limit the model size in order to avoid overfitting. Other hyper-parameters such the batch size, learning rate and dropout ratio were tuned on the development set.

Table 5: Deep learning model hyper-parameters

LSTM-FCN Parameters	Value
LSTM Hidden size	128
Number of LSTM layers	2
1D CNN channels	64,32,16
1D CNN kernel sizes	8, 5, 3
Dropout	0.2
Batch size	16
Learning rate schedule	polynomial decay
Peak learning rate	3e-4
Number of training epochs	20
Transformer Parameters	**Value**
Hidden size	128
Maximum sequence length	512
Number of layers	12
Number of attention heads	8
Feed-forward dimension	2048
Attention dropout	0.1
Activation function	GELU
Batch size	16
Learning rate schedule	polynomial decay
Peak learning rate	3e-4
Pre-training epochs	200

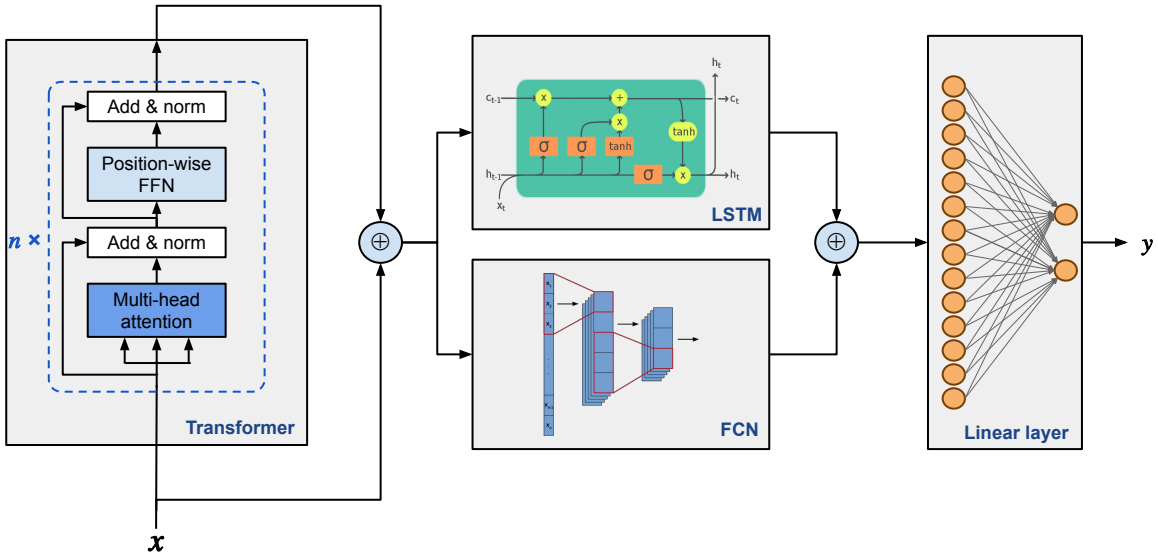

Figure 8: Larger view of the deep learning-based architecture for IMU time-series classification. $\mathbf{x}$ is the multi-variate time-series input data and y is the classification label. The $\oplus$ symbol between modules indicates concatenation.

Table 6: Cross validation results. For the feature-based setup, we report results from XGBoost model and top ranked 50 features on the validation set. The deep learning model uses the architecture in Figure 6 where the input is passed to the pre-trained Transformer without input concatenation and no fine-tuning of the Transformer. (SP: Specificity, BA: Balanced accuracy, AUC: Area under the ROC curve). The DoseMate dataset only includes data from participants of the study. Instances of open datasets are used as negative examples (no intakes) in addition to the DoseMate dataset.

	XGBoost			Deep learning		
Dataset setup	**SP**	**BA**	**AUC**	**SP**	**BA**	**AUC**
DoseMate only	.67	.82	.90	**.88**	**.90**	**.94**
DoseMate+ Cafeteria Shen et al. (2016)	.97	.97	.99	**.99**	**.99**	**1.00**
DoseMate+ Daily sports (left arm) Altun et al. (2010)	.92	.94	.98	**.96**	**.97**	**.99**
DoseMate+ Daily sports (right arm) Altun et al. (2010)	.94	.95	.98	**.97**	**.98**	**.99**
DoseMate+ FIC Kyritsis et al. (2019)	.85	.91	.95	**.96**	**.97**	**.99**
DoseMate+ Handy Açıcı et al. (2018)	.83	.89	.94	**.94**	**.95**	**.98**
DoseMate+ MHealth Banos et al. (2014)	.87	.93	.95	**.97**	**.98**	**.99**
DoseMate+ Pamap2 Reiss and Stricker (2012)	.88	.94	.94	**.96**	**.97**	**.99**
DoseMate+ IM-WSHA Tahir et al. (2020)	.78	.87	.93	**.93**	**.94**	**.97**
Average	.86	.91	.95	**.95**	**.96**	**.98**