Homekit2020: A Benchmark for Time Series Classification on a Large Mobile Sensing Dataset with Laboratory Tested Ground Truth of Influenza Infections

Mike A. Merrill MIKEAM@CS.WASHINGTON.EDU
Esteban Safranchik ESTEBANS@CS.WASHINGTON.EDU
University of Washington

Arinbjörn Kolbeinsson ARINBJORN@EVIDATION.COM
Piyusha Gade PGADE@EVIDATION.COM
Ernesto Ramirez ERAMIREZ@EVIDATION.COM
Evidation Health

Ludwig Schmidt SCHMIDT@CS.WASHINGTON.EDU
University of Washington, Allen Institute for Artificial Intelligence

Luca Foschini LUCA.FOSCHINI@SAGEBASE.ORG
Sage Bionetworks

Tim Althoff ALTHOFF@CS.WASHINGTON.EDU
University of Washington

Abstract

Despite increased interest in wearables as tools for detecting various health conditions, there are not as of yet any large public benchmarks for such mobile sensing data. The few datasets that *are* available do not contain data from more than dozens of individuals, do not contain high-resolution raw data or do not include dataloaders for easy integration into machine learning pipelines. Here, we present Homekit2020: the first large-scale public benchmark for time series classification of wearable sensor data. Our dataset contains over 14 million hours of minute-level multimodal Fitbit data, symptom reports, and ground-truth laboratory PCR influenza test results, along with an evaluation framework that mimics realistic model deployments and efficiently characterizes statistical uncertainty in model selection in the presence of extreme class imbalance. Furthermore, we implement and evaluate nine neural and non-neural time series classification models on our benchmark across 450 total training runs in order to establish state of the art performance.

Data and Code Availability This paper uses data from the Homekit2020 Flu Study, which is available on Synapse. We make all code used in this paper (including training scripts, data processing, and model hyperparameter configurations) available at this GitHub repository.

Institutional Review Board (IRB) The study that collected the data presented here was approved by the Western Institutional Review Board (WIRB, Puyallup, WA, USA) and the University of Washington IRB (Study #1271380).

1. Introduction

In the wake of the COVID-19 pandemic, there has been increased interest in using time series sensor data from wearables to detect respiratory viral infections (Alavi et al., 2022; Ates et al., 2021; Conroy et al., 2021; Kolbeinsson et al., 2021; Liu et al., 2022; Mishra et al., 2020; Natarajan et al., 2020; Quer et al., 2022; Föll et al., 2022; Mason et al., 2022; Grzesiak et al., 2021; Merrill and Althoff, 2023). If successful, this technology could enable high-frequency monitoring of vulnerable populations and warn users of potential infections before they transmit a virus to others. However, despite the significant interest in and high potential impact of this research, to date, there are no standardized tasks, evaluations, code packages, or benchmark datasets for this domain

(Nestor et al., 2021; Xu et al., 2021). As a result, it is typically impossible to compare methods or performance across publications, leading to duplicated efforts and little to no community consensus on how models should be evaluated.

There are multiple reasons for the scarcity of available wearable datasets. First, these datasets can be expensive and time-consuming to collect. Unlike in other domains of machine learning such as natural language processing and computer vision, where ground-truth labels can often be generated after corpus creation through post-hoc annotation, there is no way to recover ground-truth health information from raw time series data (e.g. a positive test for influenza). Therefore, labels must be collected along with raw data (e.g. with self-administered home test kits, costly lab tests, or surveys). Furthermore, wearable data and associated health labels constitute sensitive protected health information, which means that data can only be shared publicly with the informed consent of study participants. This has prevented many research groups from sharing their datasets.

Those datasets which *are* public are typically quite small (e.g. fewer than forty infected participants; (Mishra et al., 2020)) or only contain manually featurized and highly aggregated data that are not suitable as inputs for training deep learning models (Wang et al., 2016; Xu et al., 2022). Furthermore, these data are not preprocessed for deep learning tasks (Alavi et al., 2022) (e.g. windowed, missing values filled, reproducibly split into train/test sets). Collectively, these limitations present a significant barrier to entry for new researchers in machine learning for wearables, who must obtain IRB approval, recruit and manage participants, collect and clean data, and develop their own data processing pipelines. It is commonly reported for research groups to spend years preparing data before they can evaluate their first model (Xu et al., 2021).

Furthermore, most mobile sensing applications involve datasets that are markedly different from existing time series classification benchmarks. A popular recent time series classification benchmark is the UEA multivariate time series classification archive (Bagnall et al., 2018). Unlike the high-resolution time series relevant to mobile sensing (e.g. more than 10,000 observations for even a single week of minute-by-minute data), datasets in this benchmark are significantly shorter (fewer than 2,000 observations). Further, none of them contain any examples with missing data, which is endemic in virtually all wearable datasets. As a result, it is not known how well time series classification methods transfer to mobile sensing applications.

In this paper, we address the aforementioned challenges by sharing the Homekit2020 benchmark [1], a publicly available collection of 592,000 days of minute-level-resolution Fitbit data across 3 channels (sleep, heart rate, and step count) from 5,196 participants, combined with high-quality PCR assay test results for influenza. We describe the study protocol and data processing, and provide instructions for accessing the data (Section 3). This dataset is ten times larger than the largest public wearable device dataset to date (by 4237 participants) and additionally includes gold-standard laboratory tests instead of subjective self-reports. We go on to formalize two evaluation methods for cross-validating multivariate time series models on these data (Section 4.1) and provide a series of benchmark tasks for behavioral modeling in the context of influenza detection (Section 4.2). Finally, we evaluate several neural and non-neural baselines on our benchmark in order to establish baseline performance on this benchmark by evaluating SOTA models from influenza detection and time series classification (Section 5). We make code and detailed instructions for obtaining data available at `https://github.com/behavioral-data/Homekit2020`.

2. Related Work

Detecting Viral Infections with Wearables. There is a significant body of recent work on detecting respiratory viral infections with data from wearables (Alavi et al., 2022; Ates et al., 2021; Conroy et al., 2021; Kolbeinsson et al., 2021; Liu et al., 2022; Mishra et al., 2020; Natarajan et al., 2020; Quer et al., 2022; Föll et al., 2022; Mason et al., 2022; Grzesiak et al., 2021; Merrill and Althoff, 2023). Notably, each publication uses unique datasets, tasks, and evaluation metrics, further underscoring the need for a consistent public benchmark to facilitate comparisons between models. Methods range from non-neural models like XGBoost trained on day-level data (Grzesiak et al., 2021) to complex neural methods trained on raw minute-level-resolution sensor data (Merrill and

1. A non-public version of this dataset was previously used in a paper focused on pretraining and transfer learning methods (Merrill and Althoff, 2023). Here, we share these data publicly with additional baselines, tasks, a code library, an expanded evaluation framework, and detailed documentation.

Althoff, 2023). Furthermore, these sensitive human subjects' data are not typically shared with the public. We note that while Kolbeinsson et al. (2021) uses the dataset presented in this paper, it does not make it public and focuses instead on distinct pretraining and downstream tasks rather than benchmarking.

Existing Datasets. The closest public dataset is a collection of raw wearable data and COVID-19 diagnoses provided by Mishra et al. (2020). In comparison with Homekit2020, this dataset uses self-reported diagnoses rather than ground-truth PCR assays, has substantially fewer infected individuals (32 versus 206), and is not packaged with a pip-installable evaluation toolkit that allows ML researchers to rapidly experiment. Another related dataset is GLOBEM from Xu et al. (2022), which provides a multi-year aggregated dataset of phone sensor readings from 497 users and participant responses to surveys about their mental health. In comparison, our dataset contains roughly ten times the number of unique users, and three times the total quantity of data (Table 1). The CrossCheck dataset from Wang et al. (2016) contains smartphone sensor data and labels for schizophrenic relapse and relevant symptoms, but only provides data from 36 patients, and all data is aggregated at the hour level.

Bagnall et al. (2018) provide the most popular benchmark for multivariate time series classification models, but none of the component datasets match the characteristics of wearable data. Specifically, the time series in Bagnall et al. (2018) are shorter (fewer than 2,000 observations in length), do not contain missing data (which are quite common in sensing applications), and do not exhibit significant class imbalance (compared to the substantial imbalance in realistic tasks for wearable data (Section 4.3)).

3. The Homekit 2020 Dataset and Toolkit

3.1. Study Description

Homekit2020 is a 4-month prospective decentralized study run on the Evidation Studies platform (Kotnik et al., 2022). The aim of the study is to understand if data from consumer wearables and self-reported symptoms can be used to detect the onset of a respiratory illness.

Feature	Description
# of participants	5034
# of participants who tested flu positive	206
Mean number of days of data	114
Mean % of missing data per day (±SD)	9.8% (21%)
Daily Questionnaire completion rate	85%
Mean age (±SD)	37.7 (10.2)
% female	72%
Mean BMI (±SD)	30.3 (20.3)
# of US States Represented	50
% White participants	94.1%
% Black participants	4.6%
% Asian participants	4.2%

Table 1: Summary statistics for the Homekit2020 Flu Monitoring Study (Section 3)

3.2. Study Demographics and Statistics

The study involved 5,034 participants, who were recruited from the Evidation platform, targeting adults (age $\geq$ 18 years) residing in the United States with an active Fitbit wearable sensor connection. Study enrollment began in December 2019, shortly after the 2019-2020 influenza season began. All eligible participants owned a wearable Fitbit device capable of capturing steps, sleep and heart rate data, and agreed to wear the device as much as possible for the duration of the study. Dataset statistics are summarized in Table 1. Overall, the fraction of missing data was low (9.8%, translating to an average of 21.6 hours of data per day) and the Daily Questionnaire completion rate was high (85%). While this study represents the largest public dataset of its kind, and participants were spread across 50 US states, limitations include that the study sample skewed white and female. Often, mobile health datasets include largely very healthy individuals, limiting the representation of and value to the broader population. However, in our dataset, the mean BMI is 30.3, which is similar to the average BMI in the U.S. population (29.1 for men, and 29.6 for women).

3.3. Study Flow

During the enrollment period, participants signed an electronic informed consent, completed a baseline survey, and activated and connected their wearable Fitbit devices to the Evidation studies platform. Over the following 4 months of the study (120 days), each participant was sent a daily survey that asked about the presence of any influenza-like illness (ILI) symptoms in the past 24 hours. Participants who indicated they had experienced ILI symptoms in the

previous 24 hours were given the Daily Follow-up A survey, which included items to assess symptom onset and severity, and were directed to open their flu@home test kit and self-administer the flu@home test. At the lab, PCR testing was performed to detect the presence of different types of respiratory viruses. If participants reported not experiencing ILI symptoms over the past 24 hours or were recovering from severe ILI symptoms, they were given the Daily Follow-up B survey which asked about quality of life indicators and, if applicable, about residual ILI symptom severity upon recovery.

Participants who completed the flu@home test were also asked to complete a follow-up survey (Recovery & flu@home Experience Survey) 14 days later, which asked questions about their current health status including potential recovery. Fitbit devices were worn for the entirety of the study, and for an additional month after the end of the 120-day survey period. The devices captured data at minute-level granularity for steps, sleep and heart rate. The data were also aggregated to day-level features as shown in Table 2.

An overview of the study design for data collection is shown in Figure 1. More details about the dataset, study design, data collection, demographics and potential biases can be found in our Appendix.

3.4. Raw Data Description

The raw data consists of a) the three wearable sensor channels: heart rate, steps and sleep, all at the minute-level resolution, b) a comprehensive initial questionnaire c) daily surveys on ILI symptoms and d) results from the PCR diagnostic tests including the type of virus detected.

3.5. Data Processing

As part of this data release, we provide versions of the raw data that are preprocessed and stored for compatibility with Pytorch dataloaders, as well as a set of day-level manually defined features.

3.5.1. Minute Level Data

Some processing is necessary before these data can be passed to a model. We take the following steps to prepare the raw signals for our experiments:

1. The sleep state data stream is split into three binary signals which indicate if the study participant is in light sleep, deep sleep, or awake.

2. The heart rate, step count, and sleep state streams are resampled to a period of one minute. Within each one-minute window, these streams are aggregated by their mean, sum, and maximum respectively. Missing data are filled with zeros.

3. Three binary channels - one for each data stream - are added to the time series to indicate if data is missing within each period.

4. The data are grouped into one-week rolling windows and saved in the Petastorm format, which extends Apache Parquet to natively support multidimensional arrays and enable fast data loading into Pytorch via Spark.

3.5.2. Hour Level Data

Some models are not designed to accept inputs as large as one week's worth of minute-level raw wearable data (Section 4.4), such as transformer models that scale quadratically with the input length. Accordingly, we also provide a dataset that is identical to the minute-level data (Section 3.5.1) but is subsampled at the hour level.

3.5.3. Day Level Data

Many contemporary models for the classification of time series data from wearables rely on manually defined features (Zhang et al., 2021; Nair et al., 2019; Lin et al., 2020; Hafiz et al., 2020; Buda et al., 2021; Mairittha et al., 2021; Meegahapola et al., 2021). In order to facilitate the comparison of these models with neural models that operate on raw data, we provide a set of features (queried through the Fitbit API) that are calculated for every user every day (Table 2).

3.6. Data Availability

All data, including raw data, daily surveys, lab test results, and processed data (Section 3.5) are available through Synapse at (link withheld to protect anonymity). Note that there is no easily identifiable information in this dataset and that all study participants in this dataset have consented to their data being shared. Due to the sensitive nature of these data, researchers must verify their identities through Synapse, submit a brief (1-3 paragraph) research

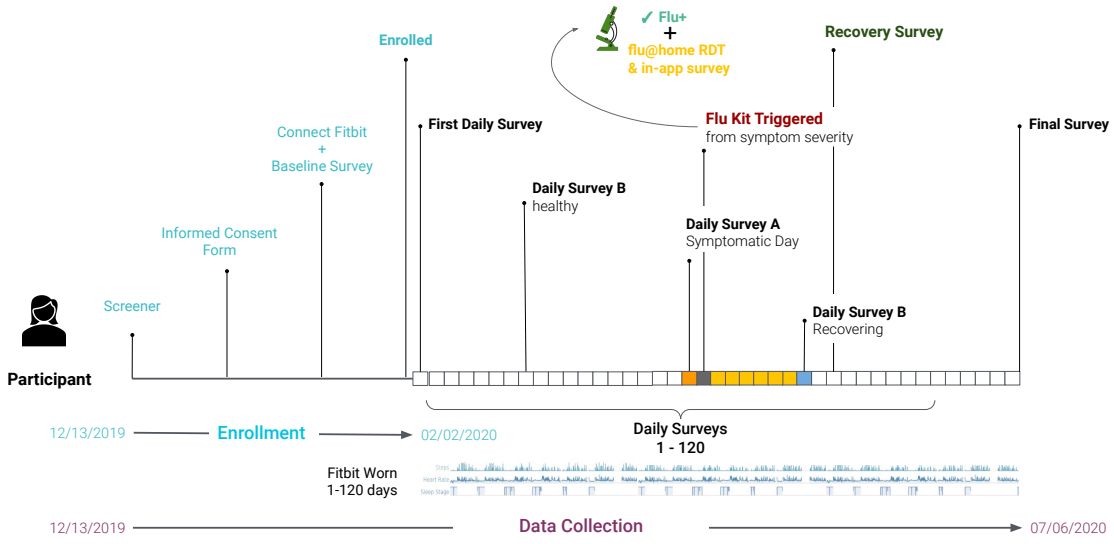


Figure 1: An overview of the study design used for Homekit2020. If a participant reported symptoms on a given day (highlighted in orange) they were sent a test-kit the following day (grey). The symptomatic period (yellow) lasted until the participant reported no symptoms (blue). (Section 3.1)

Feature Name	Feature Description
Resting heart rate (HR)	Avg. HR while still
Main minutes in bed	Longest period in bed
Sleep efficiency	Time sleeping over time in bed
Nap count	Number of naps
Total asleep minutes	Total time spent sleeping
Total in bed minutes	Total time spent in bed
Active calories	Calories burned from exercise
Calories out	Total calories burned
Base metabolic rate	Calories passively burned
Sedentary minutes	Time spent not moving
Lightly active minutes	Time spent lightly active
Fairly active minutes	Time spent lightly exercising
Very active minutes	Time spent actively exercising
Missing HR	Indicates missing HR data
Missing sleep	Indicates missing sleep data
Missing steps	Indicates missing steps
Missing day	Indicates missing all data

Table 2: Summary of day level features (Section 3.5.3), calculated for every user and on each day. "Missing" features are binary variables which are 1 if more than one hour of data is missing, and 0 otherwise.

plan, and accept certain terms and conditions including an agreement to never attempt to de-identify the data.

We also provide detailed instructions for fetching and installing the data via the Synapse API in this project's main GitHub repository: https://github.com/behavioral-data/Homekit2020.

3.7. Adding New Models

We provide preprocessed data, evaluation tasks, and predefined metrics to make it possible to evaluate new models on these data in fewer than a dozen lines of code (see GitHub for examples). The repository is structured as a Python library and can be installed with a single pip command. All code is available under the MIT License.

4. Benchmark Design

Here we provide an overview of the Homekit2020 benchmark for influenza detection.

4.1. Cross-Validation Schemes

It can be difficult to define tasks that faithfully replicate real-world conditions. Frequently, mobile sensing models:

- train on data from the future (e.g. by training on a user's data from the end of the collection period and evaluating on data from the beginning (Wang et al., 2016)).

- use data collected in laboratory settings with limited ecological validity (e.g. collecting voice samples from COVID-19 patients in a lab (Ismail et al., 2020)).
- make predictions only if a user supplies sufficient data by using a device frequently or regularly responding to surveys (e.g. by filtering users by study compliance (Malik et al., 2020; Merrill and Althoff, 2023; Wang et al., 2014)).

These practices may overestimate performance in diagnostic settings where a model would only have access to data from the past, rely on in-situ data, and would be most useful if it could function even in the presence of commonly missing data, including surveys (Nestor et al., 2021; Ismail et al., 2020). Therefore, we formally define two cross-validation schemes inspired by "real-world" deployment scenarios (Figure 2)

Temporal Split. As first proposed in Merrill and Althoff (2023), we structure our prediction tasks to emulate the following realistic scenario:

> Given training data from the first half of a flu season, how well can a model predict symptoms and infections in the second half of the flu season for every user on every day?

This scenario is based around surveillance testing, where a population is frequently tested and positive individuals are asked to undertake additional testing or self-isolate (Mercer and Salit, 2021; Merrill and Althoff, 2023). Results on tasks with this data split can be used to assess a model's ability to generalize with respect to distributional shifts over time (e.g. seasonal variations in behavior such as spending more time inside during the early winter than the spring).

User Split. For this set of prediction tasks the model is trained on data from one subset of participants and is tested on data from a distinct subset of participants. Its premise is the following scenario:

> Given training data from one randomly selected subset of study participants, how well can a model predict symptoms and infections in a distinct subset for every user on every day?

We note that it would be desirable to evaluate a model on a test set with users who were *both* unseen *and* temporally separated from training data. However, this would limit us to training on half as much data and would not produce results that are directly comparable to those we report here.

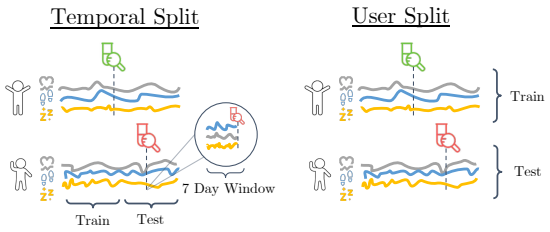


Figure 2: We evaluate models with two cross-validation schemes which are designed to reflect real world conditions (Section 4.1).

Results on tasks with this data split can be used to assess a model's ability to generalize to new users.

Additionally, our tasks only use data from the seven days prior to a predicted event so that no information from the future informs a prediction about the past. We also include no explicit information about a user's identity (e.g. participant id or demographics) to encourage models to learn generalizable motifs about activity data rather than facets of individual users' behavior. This evaluation setting follows existing best-practice recommendations and avoids falsely overstating the level of performance (Nestor et al., 2021).

4.2. Tasks

We evaluate methods on the five behavioral modeling tasks from Merrill and Althoff (2023) using a rolling seven day window of data for each prediction (Figure 2):

- **Flu Positivity:** Will the participant produce a nasal swab that tests positive for the flu today? This task emulates existing surveillance studies for both flu and COVID-19 where users are frequently tested for respiratory viral infection (Chu et al., 2020; Fusco et al., 2020).
- **Severe Fever:** Will the participant report a severe fever (defined as three or more on a four-point Likert scale) today?
- **Severe Cough:** Will the participant report a severe cough (defined as three or more on a four-point Likert scale) today?
- **Severe Fatigue:** Will the participant report severe fatigue (defined as three or more on a four-point Likert scale) today?

- **Flu Symptoms:** Will the participant report two or more flu symptoms (including cough, fever, and fatigue) of *any* severity today? This prediction is important because preliminary screening for flu typically recommends a patient for additional treatment or testing if they report some combination of two or more symptoms (CDC, 2021b), and this was the criterion used in the flu monitoring study that produced the evaluation dataset as well.

4.3. Class Imbalance Presents a Problem for Comparing Model Performance

Extreme class imbalance makes rigorously comparing model performance on individual tasks challenging, as it leads to large confidence intervals across many common test statistics for evaluation metrics. For example, under the DeLong test, a common test for comparing the ROC AUC of two classifiers, the variance of the difference in AUCs is proportional to $\frac{1}{(N-m)m}$, where N is the size of the dataset and m is the number of true positive examples (DeLong et al., 1988). This quantity is maximized when $m = 1$, or when there is only one true positive example in the data set.

As noted in Merrill and Althoff (2023), population health datasets like this regularly exhibit extreme class imbalances since, intuitively, most people are not sick on any given day. According to the CDC, the average American has a 10% chance of a symptomatic flu infection in a 365-day period (CDC, 2021a). This corresponds to a 1:3,650 class imbalance, similar to the 1:2,760 ratio in our dataset.

4.3.1. Model Selection on Individual Tasks

Stochastic optimizers induce training variance, therefore making it difficult to compare the performance across methods. This problem is exacerbated by class imbalance He and Garcia (2009). Frequently, machine learning benchmarks report the average and standard error scores across n randomly-seeded model runs (Dodge et al., 2020; Colas et al., 2018). These statistics can be used to construct a confidence interval (CI) of a model's performance with a t-distribution. However, such confidence estimates unrealistically assume normally-distributed test scores and provide unreliable error bounds for small values of n.

Input: model list $F = [f_1, ..., f_n]$, evaluation function E, number of bootstraps N, test data $\mathcal{D}$
Output: 95% CI of f evaluated with M on $\mathcal{D}$
$S \leftarrow$ empty list
for $i = 1$ **to** N **do**
  $F' \leftarrow$ sample with replacement from F
  $D' \leftarrow$ sample with replacement from D
  ModelScores $\leftarrow$ empty list
  for $f' \in F'$ **do**
    append $E(f', \mathcal{D}')$ to ModelScore
  end for
  append mean of ModelScores to S
end for
lower bound $\leftarrow$ 2.5-th percentile of S
upper bound $\leftarrow$ 97.5-th percentile of S

Algorithm 1: Hierarchical bootstrapping of a list of models (Section 4.3.1)

To ensure sound evaluation of the reported baselines, we obtain nonparametric estimates of the average performance of $n = 5$ models with 200 hierarchical bootstraps Ren et al. (2010). This approach relaxes the normal assumption of t-distribution, and as illustrated in figure 3, produces a tighter bound compared to independently averaging the bootstrapped CIs of the $n = 5$ models. Algorithm 4.3.1 summarizes our approach to obtain a 95% CI.

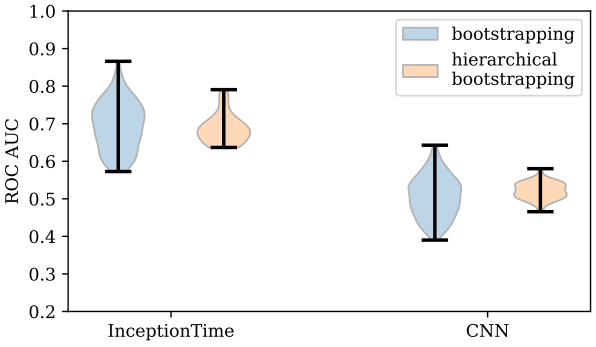


Figure 3: Confidence intervals estimated using different strategies on the Flu Positivity task.

4.3.2. Model Selection Across Tasks

Often when selecting a machine learning classifier we are interested not only in how it may perform on a given task but also in how it should perform on an arbitrary, possibly new, task. We employ critical difference plots (Brazdil and Soares, 2000) to aggregate

model performance across all tasks and both cross-validation splits. The plots first apply Friedman's statistic (Friedman, 1940) to test the null hypothesis that there is no difference between the relative performance of models, and then deploy pairwise significance tests (e.g. Wilcoxon signed rank) between classifiers (Figure 4).

4.4. Models

We evaluate the following baseline models to establish the state of the art on this benchmark:

- **XGBoost (day):** How well does a strong non-neural baseline perform? While neural models have surpassed non-neural classifiers in most CV and NLP applications, XGBoost is still commonly used in many contemporary sensing studies (e.g., (Zhang et al., 2021; Nair et al., 2019; Lin et al., 2020; Hafiz et al., 2020; Buda et al., 2021; Mairittha et al., 2021; Meegahapola et al., 2021)) in part due to its ease of use. Since boosted trees expectedly do not scale well to the thousands of observations in raw time series data, we concatenate the day-level featurized data (Section 3.5.3) within each window. We also include a version of this model trained on resampled hour-level raw data **(XGBoost (hour)).**

- **CNN:** How well does a simple CNN perform on this dataset? 1D CNNs are frequently used in timeseries classification (Pyrkov et al., 2018; Kiranyaz et al., 2021), and have been applied to data from wearable devices before (Liu et al., 2022; Shen et al., 2019; Natarajan et al., 2020). We train this model on minute-level data because smaller inputs (e.g. day level (Section 3.5.3) or hour level (Section 3.5.3)) compress to a small number of spatial dimensions after only a few layers of a standard CNN.

- **Transformer Hour Level:** Transformers are increasingly seen as strong models for a variety of tasks across many data modalities (e.g. text (Devlin et al., 2018), images (Dosovitskiy et al., 2021), and audio (Liu et al., 2020)). One challenge of training these models on raw time series data is their quadratic memory complexity with respect to input length. This makes it computationally infeasible to train such a model on high-resolution raw data (e.g. a week-long window with 10,080 observations) (Beltagy et al., 2020). To compensate for this limitation, we train this model on the resampled hour-level data (Section 3.5.2) which has a comparatively small dimensionality (168 observations in the case of a one-week window).

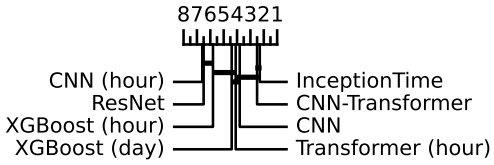
- **CNN-Transformer:** By applying a series of transformer blocks to the CNN output, a model may learn relationships between the spatial features provided by its CNN encoder (Merrill and Althoff, 2023).

- **ResNet:** While ResNet is generally considered to be far from the state of the art in the computer vision community, it is still viewed as a competitive model for multivariate time series classification (He et al., 2015). For example, it is the highest-ranking neural model on the UEA multivariate time series classification archive (Ruiz et al., 2021). We additionally include a version of ResNet trained on the resampled hour-level data (**ResNet (hour)**).

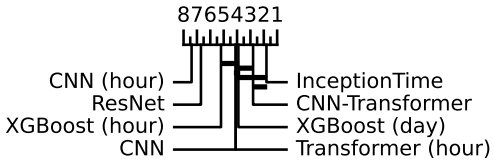
- **InceptionTime:** How well does a state of the art time series classification model perform on this task? InceptionTime (Ismail Fawaz et al., 2020)) uses an ensemble of convolutions at different temporal resolutions to extract relevant features. It is the top-performing neural model in Ruiz et al. (2021)'s comparison of multivariate time series classification models.

4.5. Metrics

We use three methods to evaluate models: ROC AUC, Precision-Recall AUC (PR AUC), and Precision at k%. **ROC AUC** is useful because its magnitude is not dependent on class balance, allowing us to compare performance across different tasks. Conversely, **PR AUC** reflects a model's ability to correctly retrieve positive examples and is impacted by class balance. In some applications, such as allocating scarce antivirals, allocating additional testing, or selecting clinical trial participants selecting the most at-risk members of a population may be more important than identifying *all* diseased individuals. **Precision at k%** allows us to measure performance among a model's most confident predictions on a test set of N examples: Precision @ $k\% = \frac{1}{n}\sum_{i=1}^{n} y_i$ where $n = \lfloor k\% \text{ of } N \rfloor$ and y_i is the list of N ground truth binary labels in decreasing order of the model's confidence.



(*a*) Critical Difference w.r.t. PR AUC



(*b*) Critical Difference w.r.t. ROC AUC

Figure 4: Critical difference diagrams with respect to precision-recall AUC and ROC AUC. Numbers indicate each model's average ranking on tasks (Section 4.2), while the thick dark line connects models which are not significantly different from one another.

5. Results

The Homekit2020 benchmark shows that training on high-frequency time series improves model performance, that there is a 100x lift above random performance within highly-confident model predictions, and that classifying data from new users is roughly as challenging as modeling data from the future.

High-frequency data can outperform XGBoost trained on coarse data. As mentioned in Section 4.3, the substantial class imbalance inherent in these data limits comparisons between classifiers on individual tasks. However, we find that InceptionTime (trained on minute-level data)(Ismail Fawaz et al., 2020) ranks first across most tasks with respect to ROC AUC and first across four tasks in PR AUC (Table 3). Notably, it significantly outranks ResNet, CNN, XGBoost (hour) and CNN (hour). (Figure 4(*a*)) with respect to PR AUC. There is less separability between models with respect to ROC (although InceptionTime still performs best on most tasks).

In Figure 5 we report model performance with hierarchically bootstrapped confidence intervals (Section 4.3.1). Notably, despite its popularity in mobile sensing literature (Section 2), XGBoost is outperformed by a neural method in all but two cases. This indicates that neural models trained on raw time series generally perform better than non-neural models trained on aggregated time series data. Overall these are very challenging tasks with limited performance. There is potential for these models to be useful, but future work should establish "how good is good enough" as current methods may not be better than symptomatic screening.

Precision on confident test examples indicates substantial lift over random baseline. In some epidemiological settings (e.g. allocating scarce antivirals or selecting clinical trial participants) identifying the most at-risk individuals in a population may be more important than identifying *all* diseased individuals. In Figure 6, we focus on precision among the model's most confident predictions (as determined by the softmax output of each model) using Precision at $k\%$. We compare model performance to a random baseline (equal to the condition's prevalence rate) and an "oracle" model which is able to perfectly rank test examples (Figure 6) While there is still substantial room for improvement at higher k (as shown by decreasing model performance), some models show significant lift over the random baseline on many tasks. For example, the CNN-Transformer achieves 3.9% precision at $k\%$=0.05%, a 100x lift over the baseline prevalence of 0.036%.

Classifying data from new users is roughly as hard as classifying data from the future. It is notable that model ROC AUC performance is comparable between the temporal and user splits. This indicates that learning to classify data "from the future" (as in the temporally split task) is roughly as difficult as learning to make predictions about a different subset of users (as in the user split task). Future work could characterize this relationship by studying changes in performance with respect to data quality, temporal distance between train and test points, and demographic similarity between users.

6. Limitations

The Homekit 2020 Benchmark provides the largest mobile sensing dataset for infectious disease detection, a set of tasks for rigorously evaluating time series classification models, and a pip-installable toolkit for running experiments on the dataset. However, our benchmark is not without limitations. Study participants were recruited across 50 U.S. states, but were

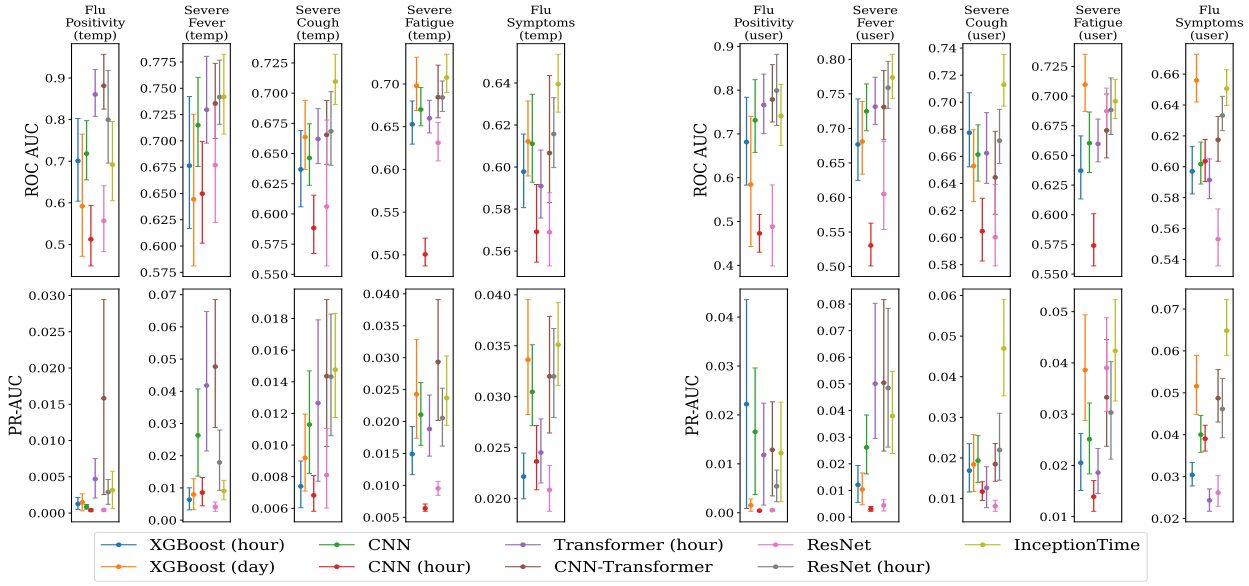


Figure 5: Model performance across all cross-validation splits (Section 4.1) and tasks (Section 4.2). Confidence intervals are calculated with hierarchical bootstrapping (Section 4.3.1)

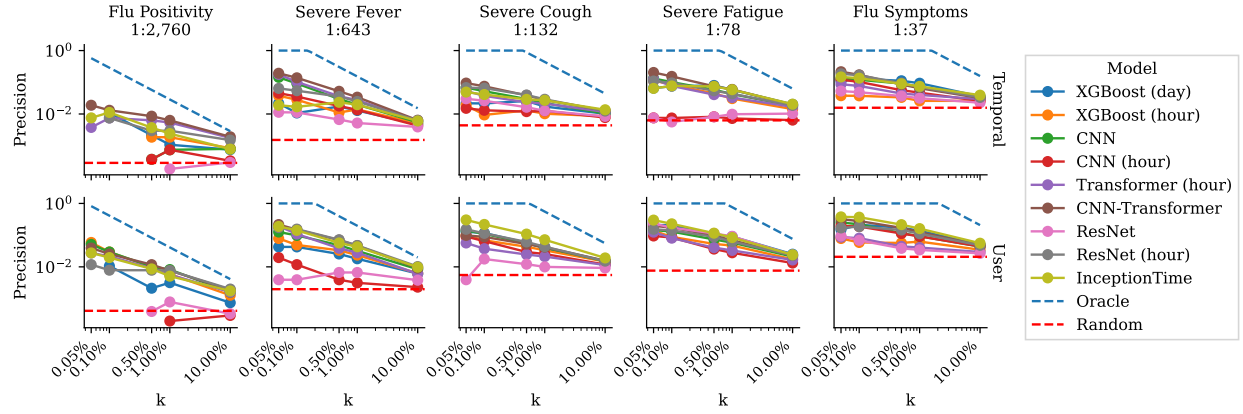


Figure 6: Precision for each model and each task, given that the model is asked to retrieve its $k\%$ most confident test set examples as positives (Section 5).

ROC AUC	Flu Positivity		Severe Fever		Severe Cough		Severe Fatigue		Flu Symptoms	
	Temp.	User	Temp.	User	Temp.	User	Temp.	User	Temp.	User
XGBoost (day)	0.566	0.561	0.638	0.686	0.667	0.656	0.693	**0.709**	0.616	**0.652**
XGBoost (hour)	0.683	0.679	0.680	0.670	0.640	0.677	0.652	0.639	0.599	0.596
CNN	0.716	0.732	0.715	0.717	0.647	0.663	0.677	0.655	0.613	0.603
CNN (hour)	0.497	0.480	0.648	0.534	0.590	0.604	0.501	0.576	0.563	0.603
Transformer (hour)	0.856	0.761	0.721	0.731	0.662	0.666	0.659	0.654	0.588	0.593
CNN-Transformer	**0.876**	0.795	0.737	0.734	0.669	0.646	0.685	0.673	0.608	0.617
ResNet	0.548	0.486	0.671	0.614	0.622	0.601	0.631	0.686	0.568	0.554
ResNet (hour)	0.794	**0.812**	**0.746**	0.762	0.669	0.670	0.682	0.694	0.618	0.633
InceptionTime	0.709	0.735	0.742	**0.779**	**0.712**	**0.714**	**0.707**	0.693	**0.641**	0.650

PR AUC	Flu Positivity		Severe Fever		Severe Cough		Severe Fatigue		Flu Symptoms	
	Temp.	User	Temp.	User	Temp.	User	Temp.	User	Temp.	User
XGBoost (day)	0.001	0.001	0.007	0.011	0.009	0.017	0.024	0.037	0.034	0.050
XGBoost (hour)	0.001	**0.019**	0.007	0.009	0.007	0.017	0.016	0.020	0.022	0.030
CNN	0.001	0.013	0.027	0.027	0.012	0.021	0.023	0.024	0.031	0.040
CNN (hour)	0.000	0.000	0.010	0.003	0.007	0.012	0.006	0.014	0.023	0.039
Transformer (hour)	0.005	0.011	0.036	**0.057**	0.010	0.013	0.019	0.019	0.024	0.025
CNN-Transformer	**0.012**	0.013	**0.046**	0.047	0.015	0.019	**0.030**	0.035	0.032	0.049
ResNet	0.000	0.001	0.004	0.005	0.009	0.008	0.010	**0.040**	0.021	0.026
ResNet (hour)	0.002	0.005	0.018	0.046	0.013	0.022	0.021	0.033	0.033	0.047
InceptionTime	0.003	0.010	0.009	0.044	**0.015**	**0.046**	0.024	0.038	**0.036**	**0.065**

Table 3: Hierarchically Bootstrapped ROC/PR AUC across all tasks. "Temp." indicates temporal split, while "User" indicates user split (Section 4.1). InceptionTime Ismail Fawaz et al. (2020) consistently performs best.

disproportionately white and female (Table 1). While a user's skin color has been shown to not be a significant source of bias in photoplethysmography (PPG) heart rate measurements (Bent et al., 2020), we encourage users of this benchmark to evaluate sensing technologies on a diverse pool of participants wherever possible. Future data collection should place an even greater emphasis on collecting data from a representative population sample.

The authors emphasize that scores on this benchmark may not be indicative of real-world performance. This benchmark contains data from one brand of wearable (Fitbit), one flu season, and a non-representative sample of the population. Method tested against this benchmark should undergo testing through clinical trials before large-scale deployment, e.g. as a feature in a commercial wearable.

7. Conclusion

We propose the Homekit2020 Benchmark and provide models, data processing code, and 14 million hours of wearable data with high-quality laboratory PCR results for influenza. Our hope is that this benchmark will provide a test bed for machine learning on wearable data for health, and accelerate progress in this important research area. Our results highlight future opportunities for modeling these data, including performance improvements from modeling high-frequency raw wearable data with neural methods. We believe that this benchmark can be used to study and evaluate machine learning methods for modeling behavioral and health data, including self-supervision, transfer learning and few/zero-shot learning methods.

Acknowledgments

The authors thank the HomeKit2020 team, including Matthew Thompson and Barry Lutz, for sharing their dataset to enable this prediction benchmark. The Homekit2020 dataset was collected through a 4-month prospective decentralized study run on the Evidation Studies platform (Kotnik et al., 2022) (Evidation Inc., San Mateo, CA). The data collection was supported by the Biomedical Advanced Research and Development Authority (BARDA Contract Number 75A50119C00036) and Audere. This research was supported in part by the Bill & Melinda Gates Foundation (INV-004841), NSF CAREER IIS-2142794, NSF grant IIS-1901386, NSF grant CNS-2025022, the Office for Naval Research (#N00014-21-1-2154), and a Microsoft AI for Accessibility grant.

References

Arash Alavi, Gireesh K. Bogu, Meng Wang, Ekanath Srihari Rangan, Andrew W. Brooks, Qiwen Wang, Emily Higgs, Alessandra Celli, Tejaswini Mishra, Ahmed A. Metwally, Kexin Cha, Peter Knowles, Amir A. Alavi, Rajat Bhasin, Shrinivas Panchamukhi, Diego Celis, Tagore Aditya, Alexander Honkala, Benjamin Rolnik, Erika Hunting, Orit Dagan-Rosenfeld, Arshdeep Chauhan, Jessi W. Li, Caroline Bejikian, Vandhana Krishnan, Lettie McGuire, Xiao Li, Amir Bahmani, and Michael P. Snyder. Real-time alerting system for COVID-19 and other stress events using wearable data. *Nature Medicine*, 28(1):175–184, January 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01593-2. URL https://www.nature.com/articles/s41591-021-01593-2. Number: 1 Publisher: Nature Publishing Group.

H. Ceren Ates, Ali K. Yetisen, Firat Güder, and Can Dincer. Wearable devices for the detection of COVID-19. *Nature Electronics*, 4(1):13–14, January 2021. ISSN 2520-1131. doi: 10.1038/s41928-020-00533-1. URL https://www.nature.com/articles/s41928-020-00533-1. Number: 1 Publisher: Nature Publishing Group.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv:1811.00075 [cs, stat]*, October 2018. URL http://arxiv.org/abs/1811.00075. arXiv: 1811.00075.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*, 2020.

Brinnae Bent, Benjamin A. Goldstein, Warren A. Kibbe, and Jessilyn P. Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine*, 3(1):18, December 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0226-6. URL http://www.nature.com/articles/s41746-020-0226-6.

Pavel B. Brazdil and Carlos Soares. A Comparison of Ranking Methods for Classification Algorithm Selection. In Jaime G. Carbonell, Jörg Siekmann, G. Goos, J. Hartmanis, J. van Leeuwen, Ramon López de Mántaras, and Enric Plaza, editors, *Machine Learning: ECML 2000*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

Teodora Sandra Buda, Mohammed Khwaja, and Aleksandar Matic. Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021. ISSN 2474-9567.

CDC. Estimated Flu-Related Illnesses, Medical visits, Hospitalizations, and Deaths in the United States — 2018–2019 Flu Season | CDC, 2021a.

CDC. Influenza Signs and Symptoms and the Role of Laboratory Diagnostics, 2021b.

CDC. CDC Museum COVID-19 Timeline, 2023. URL https://www.cdc.gov/museum/timeline/covid19.html.

Helen Y. Chu, Janet A. Englund, Lea M. Starita, Michael Famulare, Elisabeth Brandstetter, Deborah A. Nickerson, Mark J. Rieder, Amanda Adler, Kirsten Lacombe, Ashley E. Kim, Chelsey Graham, Jennifer Logue, Caitlin R. Wolf, Jessica Heimonen, Denise J. McCulloch, Peter D. Han, Thomas R. Sibley, Jover Lee, Misja Ilcisin, Kairsten Fay, Roy Burstein, Beth Martin, Christina M. Lockwood, Matthew Thompson, Barry Lutz, Michael Jackson, James P. Hughes, Michael Boeckh, Jay Shendure, and Trevor Bedford. Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform. *New England Journal of Medicine*, 2020.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments, July 2018. URL http://arxiv.org/abs/1806.08295. arXiv:1806.08295 [cs, stat].

Bryan Conroy, Ikaro Silva, Golbarg Mehraei, Robert Damiano, Brian Gross, Emmanuele Salvati, Ting Feng, Jeffrey Schneider, Niels Olson, Anne Rizzo, Catherine Curtin, Joseph Frassica, and Daniel Mcfarlane. Real-time infection prediction with wearable physiological monitoring and AI: Aiding military workforce readiness during COVID-1. Technical report, 2021. URL https://europepmc.org/article/PPR/PPR382715. Type: article.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 1988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping, February 2020. URL http://arxiv.org/abs/2002.06305. arXiv:2002.06305 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Technical Report arXiv:2010.11929, arXiv, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs] type: article.

Milton Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. pages 86–92, 1940. ISSN 0003-4851, 2168-8990.

F. M. Fusco, M. Pisaturo, V. Iodice, R. Bellopede, O. Tambaro, G. Parrella, G. Di Flumeri, R. Viglietti, R. Pisapia, M. A. Carleo, M. Boccardi, L. Atripaldi, B. Chignoli, N. Maturo, C. Rescigno, V. Esposito, R. Dell'Aversano, V. Sangiovanni, and R. Punzi. COVID-19 among healthcare workers in a specialist infectious diseases setting in Naples, Southern Italy: results of a cross-sectional surveillance study. *Journal of Hospital Infection*, 2020. ISSN 0195-6701.

Simon Föll, Adrian Lisson, Martin Maritsch, Karsten Klingberg, Vera Lehmann, Thomas Züger, David Srivastava, Sabrina Jegerlehner, Stefan Feuerriegel, Aristomenis Exadaktylos, and Felix Wortmann. A Scalable Risk Scoring System for COVID-19 Inpatients Based on Consumer-grade Wearables: Statistical Analysis and Model Development. page 42, 2022.

Emilia Grzesiak, Brinnae Bent, Micah T. McClain, Christopher W. Woods, Ephraim L. Tsalik, Bradly P. Nicholson, Timothy Veldman, Thomas W. Burke, Zoe Gardener, Emma Bergstrom, Ronald B. Turner, Christopher Chiu, P. Murali Doraiswamy, Alfred Hero, Ricardo Henao, Geoffrey S. Ginsburg, and Jessilyn Dunn. Assessment of the Feasibility of Using Noninvasive Wearable Biometric Monitoring Sensors to Detect Influenza and the Common Cold Before Symptom Onset. *JAMA Network Open*, 4(9):e2128534, September 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.28534. URL https://doi.org/10.1001/jamanetworkopen.2021.28534.

Pegah Hafiz, Kamilla Woznica Miskowiak, Alban Maxhuni, Lars Vedel Kessing, and Jakob Eyvind Bardram. Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls. *IMWUT*, 2020.

Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1558-2191. doi: 10.1109/TKDE.2008.239. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. Technical Report arXiv:1512.03385, arXiv, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs] version: 1 type: article.

Mahmoud Al Ismail, Soham Deshmukh, and Rita Singh. Detection of COVID-19 through the analysis of vocal fold oscillations. 2020.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, Nov 2020. ISSN 1573-756X. doi: 10.1007/s10618-020-00710-y.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 2021.

Arinbjörn Kolbeinsson, Piyusha Gade, Raghu Kainkaryam, Filip Jankovic, and Luca Foschini. Self-supervision of wearable sensors time-series data for influenza detection. *arXiv:2112.13755 [cs]*, December 2021. URL http://arxiv.org/abs/2112.13755. arXiv: 2112.13755.

Jack Henry Kotnik, Shawna Cooper, Sam Smedinghoff, Piyusha Gade, Kelly Scherer, Mitchell Maier, Jessie Juusola, Ernesto Ramirez, Pejman Naraghi-Arani, Victoria Lyon, Barry Lutz, and Matthew Thompson. Flu@home: the Comparative Accuracy of an At-Home Influenza Rapid Diagnostic Test Using a Prepositioned Test Kit, Mobile App, Mail-in Reference Sample, and Symptom-Based Testing Trigger. *Journal of Clinical Microbiology*, 60(3):e02070–21, March 2022. doi: 10.1128/jcm.02070-21. URL https://journals.asm.org/doi/full/10.1128/jcm.02070-21. Publisher: American Society for Microbiology.

Zongyu Lin, Shiqing Lyu, Hancheng Cao, Fengli Xu, Yuqiong Wei, Hanan Samet, and Yong Li. HealthWalks: Sensing Fine-grained Individual Health Condition via Mobility Data. *IMWUT*, pages 1–26, 2020.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423, May 2020. doi: 10.1109/ICASSP40776.2020.9054458. ISSN: 2379-190X.

Shuo Liu, Jing Han, Estela Laporta Puyal, Spyridon Kontaxis, Shaoxiong Sun, Patrick Locatelli, Judith Dineley, Florian B. Pokorny, Gloria Dalla Costa, Letizia Leocani, Ana Isabel Guerrero, Carlos Nos, Ana Zabalza, Per Soelberg Sørensen, Mathias Buron, Melinda Magyari, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Amos A Folarin, Richard JB Dobson, Raquel Bailón, Srinivasan Vairavan, Nicholas Cummins, Vaibhav A Narayan, Matthew Hotopf, Giancarlo Comi, Björn Schuller, and RADAR-CNS Consortium. Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder. *Pattern Recognition*, 123:108403, March 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108403. URL https://www.sciencedirect.com/science/article/pii/S0031320321005793.

Nattaya Mairittha, Tittaya Mairittha, Paula Lago, and Sozo Inoue. CrowdAct: Achieving High-Quality Crowdsourced Datasets in Mobile Activity Recognition. *IMWUT*, 2021.

Momin M. Malik, Afsaneh Doryab, Michael Merrill, Jürgen Pfeffer, and Anind K. Dey. Can Smartphone Co-locations Detect Friendship? It Depends How You Model It. *arXiv:2008.02919 [cs]*, 2020.

Ashley E. Mason, Frederick M. Hecht, Shakti K. Davis, Joseph L. Natale, Wendy Hartogensis, Natalie Damaso, Kajal T. Claypool, Stephan Dilchert, Subhasis Dasgupta, Shweta Purawat, Varun K. Viswanath, Amit Klein, Anoushka Chowdhary, Sarah M. Fisher, Claudine Anglo, Karena Y. Puldon, Danou Veasna, Jenifer G. Prather, Leena S. Pandya, Lindsey M. Fox, Michael Busch, Casey Giordano, Brittany K. Mercado, Jining Song, Rafael Jaimes, Brian S. Baum, Brian A. Telfer, Casandra W. Philipson, Paula P. Collins, Adam A. Rao, Edward J. Wang, Rachel H. Bandi, Bianca J. Choe, Elissa S. Epel, Stephen K. Epstein, Joanne B. Krasnoff, Marco B. Lee, Shi-Wen Lee, Gina M. Lopez, Arpan Mehta, Laura D. Melville, Tiffany S. Moon, Lilianne R. Mujica-Parodi, Kimberly M. Noel, Michael A. Orosco, Jesse M. Rideout, Janet D. Robishaw, Robert M. Rodriguez, Kaushal H. Shah, Jonathan H. Siegal, Amarnath Gupta, Ilkay Altintas, and Benjamin L. Smarr. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study. *Scientific Reports*, 12 (1):3463, December 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07314-0. URL https://www.nature.com/articles/s41598-022-07314-0.

Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports. *IMWUT*, 2021.

Tim R. Mercer and Marc Salit. Testing at scale during the COVID-19 pandemic. *Nature Reviews Genetics*, 2021.

Mike A. Merrill and Tim Althoff. Self-supervised Pretraining and Transfer Learning Enable Flu and

COVID-19 Predictions in Small Mobile Sensing Datasets. *Conference on Health, Inference, and Learning*, 2023.

Tejaswini Mishra, Meng Wang, Ahmed A. Metwally, Gireesh K. Bogu, Andrew W. Brooks, Amir Bahmani, Arash Alavi, Alessandra Celli, Emily Higgs, Orit Dagan-Rosenfeld, Bethany Fay, Susan Kirkpatrick, Ryan Kellogg, Michelle Gibson, Tao Wang, Erika M. Hunting, Petra Mamic, Ariel B. Ganz, Benjamin Rolnik, Xiao Li, and Michael P. Snyder. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering*, 4(12):1208–1220, December 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-00640-6. URL https://www.nature.com/articles/s41551-020-00640-6.

Suraj Nair, Kiran Javkar, Jiahui Wu, and Vanessa Frias-Martinez. Understanding Cycling Trip Purpose and Route Choice Using GPS Traces and Open Data. *IMWUT*, 2019.

Aravind Natarajan, Hao-Wei Su, and Conor Heneghan. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *npj Digital Medicine*, 3(1):1–8, November 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00363-7. URL https://www.nature.com/articles/s41746-020-00363-7. Number: 1 Publisher: Nature Publishing Group.

Bret Nestor, Jaryd Hunter, Raghu Kainkaryam, Erik Drysdale, Jeffrey B Inglis, Allison Shapiro, Sujay Nagaraj, Marzyeh Ghassemi, Luca Foschini, and Anna Goldenberg. Dear watch, should i get a covid-19 test? designing deployable machine learning for wearables. *medRxiv*, 2021.

Timothy V. Pyrkov, Konstantin Slipensky, Mikhail Barg, Alexey Kondrashin, Boris Zhurov, Alexander Zenin, Mikhail Pyatnitskiy, Leonid Menshikov, Sergei Markov, and Peter O. Fedichev. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Scientific Reports*, 2018.

Giorgio Quer, Matteo Gadaleta, Jennifer M. Radin, Kristian G. Andersen, Katie Baca-Motes, Edward Ramos, Eric J. Topol, and Steven R. Steinhubl. Inter-individual variation in objective measure of reactogenicity following COVID-19 vaccination via smartwatches and fitness bands. *npj Digital Medicine*, 5(1):49, December 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00591-z. URL https://www.nature.com/articles/s41746-022-00591-z.

Shiquan Ren, Hong Lai, Wenjing Tong, Mostafa Aminzadeh, Xuezhang Hou, and Shenghan Lai. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, 37 (9):1487–1498, September 2010. ISSN 0266-4763. doi: 10.1080/02664760903046102. URL https://doi.org/10.1080/02664760903046102. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02664760903046102.

Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, March 2021. ISSN 1573-756X. doi: 10.1007/s10618-020-00727-3. URL https://doi.org/10.1007/s10618-020-00727-3.

Yichen Shen, Maxime Voisin, Alireza Aliamiri, Anand Avati, Awni Hannun, and Andrew Ng. Ambulatory Atrial Fibrillation Monitoring Using Wearable Photoplethysmography with Deep Learning. In *KDD*, 2019.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp*, 2014.

Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *UbiComp*, 2016.

Xuhai Xu, Jennifer Mankoff, and Anind K. Dey. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. *CCF Transactions on Pervasive Computing and Interaction*, 3(4):344–366, December 2021. ISSN 2524-5228. doi: 10.1007/s42486-021-00072-4. URL https://doi.org/10.1007/s42486-021-00072-4.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve Riskin, Jennifer Mankoff, and Anind K Dey. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. 2022.

Yunke Zhang, Fengli Xu, Tong Li, Vassilis Kostakos, Pan Hui, and Yong Li. Passive Health Monitoring Using Large Scale Mobility Data. *IMWUT*, 2021.

Appendix A. Adding New Models to Homekit2020

Homekit2020 is both a dataset, a Python package, and a command line utility for training neural models on sensor data. We have structured the package so that it is trivial to implement a new model by subclassing the provided model class and implementing a forward pass in PyTorch:

```
from src.models.models import ClassificationModel
import torch.nn as nn

class FluLSTM(ClassificationModel):

    def __init__(self, **kwargs) -> None:
        super().__init__(**kwargs, hidden_dim : int = 32,
                                   embedding_dim : int = 128)
        self.lstm = nn.LSTM(hidden_dim, embedding_dim)
        self.criterion = nn.CrossEntropyLoss()

    def forward(self, x,labels):
        preds = self.lstm(x)
        loss =  self.criterion(preds,labels)
        return loss, preds
```

The model can then be evaluated on any of our tasks from the command line:

```

python src/models/train.py fit --model.class_path path.to.FluLSTM\
                               --config configs/tasks/PredictFluPos.yaml
```

Appendix B. A note on the 2019-2020 Flu Season and COVID-19

Epidemiologists are still studying the early days of the COVID-19 pandemic, but CDC consensus is that significant community transmission was not prevalent in the United States until the last week of March 2020 CDC (2023). Importantly, a typical flu season peaks in February. In our dataset the last flu-positive result was logged on March 21, 2020 with a peak around February 1, 2020. These facts combined lead us to believe that while there may have been some overlap with COVID cases, the disease was likely not nearly as prevalent within our population as it would have been during, say, the Omicron wave in Summer 2021.

We should also note that while there were several high-profile early outbreaks of COVID-19 in the United States during March and April 2020 (e.g. Washington State, New York City) participants in this study were geographically dispersed throughout all 50 states (Table 1).

Nonetheless, it is possible that some self-reported symptoms reflect an early COVID-19 infection. Importantly, nowhere in the paper do we claim that symptoms are necessarily caused by any particular disease, be it influenza, RSV, or COVID-19. Symptomatic reports should be taken at face value. For example, if a study participant reports fatigue then it should be assumed that participant was fatigued and nothing more.

Appendix C. Study overview

Homekit2020 is a 4-month prospective observational decentralized cohort study run on the Evidation Studies platform (Evidation Inc., San Mateo, CA) Kotnik et al. (2022) platform. The aim of the study is to understand if data from consumer wearables and self-reported symptoms can be used to detect the onset of a respiratory illness. The study had 3 objectives, defined in the study protocol before the study started and accompanied by statistical analysis plans (SAP). The objectives were:

- To develop a database of everyday behavior data associated with participants who reported their health status and activity data over the study duration.
- To Investigate the effectiveness of using behavioral and physiological data derived from wearable devices to develop classification models for flu and non-flu respiratory viral infections (RVI) at varying levels of training label confidence.
- To develop a regression model for states with most flu cases to forecast influenza-like illness (ILI) infection rates.

The study protocol was not pre-registered on any public register.

Funding. The study was supported by the Biomedical Advanced Research and Development Authority (BARDA Contract Number 75A50119C00036) and Audere.

Ethical Review. All enrolled participants completed an online informed consent form agreeing to study protocols. The study was approved by Western Institutional Review Board, Inc. (Puyallup, WA).

C.1. Study design

Duration. Recruitment started on December 13th 2019 and completed on February 2nd 2020. Wearable data collection started on the date of recruitment. The study ran from December 2019 to June 2020 with daily surveys and minute-level wearable data collection. At the end of the study, participants were asked to complete a final survey which asked questions about their overall experience with the study including their influenza vaccination history for the 2019-2020 flu season.

Recruitment. The study involved $5,196$ participants, who were recruited from the Evidation platform, targeting adults (age $\geq$ 18 years) residing in the United States with an active Fitbit wearable sensor connection. Study enrollment began in December 2019, shortly after the 2019-2020 influenza season began. All eligible participants owned a wearable Fitbit device capable of capturing steps, sleep and heart rate data, and agreed to wear the device as much as possible for the duration of the study.

Inclusion/Exclusion criteria. The criteria for selection were:

- currently living in the United States
- ability to read, speak, and understand English
- own and regularly wear a Fitbit device that tracks steps, sleep, and heart rate
- having not been diagnosed with the flu in the 3 months before the start of the study
- willingness to complete a daily online survey for the duration of the study
- have an iPhone, iPad, or Android smartphone or tablet
- willingness to download an app if they experience flu-like symptoms
- willingness to complete an at-home flu test kit and send sample to a laboratory using a pre-paid shipping label

The criteria for exclusion were:

- diagnosed with flu by a healthcare professional in the past three months
- currently enrolled in another flu study being conducted by Evidation Health.

Demographic		All Participants (n = 5229)
Age	Mean ± SD	37.7 (10.2)
	Median	36.0 (14.0)
	Min - Max	18.0 - 79.0
bmi	Mean ± SD	30.3 (20.3)
	Median	28.5 (9.3)
	Min - Max	13.2 - 70.9
Gender, N (%)	Female	3763 (72.0%)
	Male	1453 (27.8%)
	Other/Non-binary	11 (0.2%)
Pregnant, N (%)	Yes	66 (1.3%)
	No	3710 (71.0%)
Ethnicity, N (%)	Non-Hispanic or Latino	4922 (94.1%)
	Hispanic or Latino	307 (5.9%)
Race, N (%)	White	4781 (91.4%)
	Black or African American	240 (4.6%)
	Asian	219 (4.2%)
	American Indian or Alaska Native	70 (1.3%)
	Native Hawaiian or Pacific Isl	18 (0.3%)
	Other	92 (1.8%)

Demographic		All Participants (n = 5229)
Education, N (%)	Did not complete high school, no diploma	25 (0.5%)
	High school graduate, diploma or the equivalent	299 (5.7%)
	Some college, no degree	874 (16.7%)
	Trade/technical/vocational training	219 (4.2%)
	College graduate, associate or bachelor's degree	2572 (49.2%)
	Graduate degree	1051 (20.1%)
	Doctorate degree	180 (3.4%)
	Prefer not to answer	< 10 (< 0.2%)*
Income, N (%)	Less than $25,000	361 (6.9%)
	$25,000 - $34,999	422 (8.1%)
	$35,000 - $49,999	757 (14.5%)
	$50,000 - $74,999	1148 (22.0%)
	$75,000 - $99,999	924 (17.7%)
	$100,000 - $149,000	917 (17.5%)
	$150,000 or more	385 (7.4%)
	Prefer not to answer	315 (6.0%)

Figure C.7: Participant demographics of Homekit2020. The cohort is skewed young, overweight, female, white, and college educated compared to the general US population. Overrepresented demographics are highlighted in gray.

Study Procedures. During the enrollment period, participants signed an electronic informed consent, completed a baseline survey, activated and connected their wearable Fitbit devices to the Evidation studies platform. Over the following 4 months of the study (120 days), each participant was sent a daily survey that asked about the presence of any influenza-like illness (ILI) symptoms in the past 24 hours. Participants who indicated they had experienced ILI symptoms in the previous 24 hours were given daily follow up surveys to learn more about their symptoms. If symptoms were present, they were directed to self-administered a flu test, which would give both an immediate results for generic influenza infection, and later be analyzed at a lab to determine the specific type of virus. Participants were compensated for taking part in the study.

Demographics and Representatives. The study aimed at enrolling a *convenience sample* of the population. The sample is not representative of the US population as it skews young, overweight, female, white, college educated, as noted in Figure C.7. Further studies should be aimed at correcting this bias by recruiting a more representative population.

Completion metrics. Adherence to protocol was high across the board, with minimal missing data across data sources, as shown in Figure C.8

Appendix D. Datasheet for Homekit2020

To aid reproducibility, we borrow the datasheet for datasets [2] to report details about our dataset in a standadized way.

2. Gebru, Timnit, et al. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*

#	Key Metrics	Completion Rates
1	Baseline Questionnaire completion rates	5,229 (100%)
2	Daily Questionnaire (120 days) completion rates	527,877 total (avg completion rate of 85%)
3	Total Flu Kits Triggered & Completed	Triggered: 1242 Completed: 1009 (81%)
4	Recovery Questionnaire completion rates	998 (99%)
5	Final Survey completion rates	4933 (94%)
6	Average number of days with Fitbit data (out of 120 days)	110 valid* days (92%)

* valid : >10 h of daily weartime

Figure C.8: Key completion metrics for the dataset

D.1. Dataset

- **Why was the dataset created?** This dataset was created to:
  - develop a database of everyday behavior data associated with participants who reported their health status and activity data over the study duration
  - investigate the effectiveness of using behavioral and physiological data derived from wearable devices to develop classification models for flu and non-flu RVI at varying levels of training label confidence.
  - develop a regression model for states with most flu cases to forecast ILI infe ction rates
- **Who funded the creation of the dataset?** The study was supported by the Biomedical Advanced Research and Development Authority (BARDA) and Audere.
- **What are the instances? How many instances of each type are there?** Instances consist of the 5,196 participants who took part in the study, of which 5,034 consented to data sharing. Of those, 1,001 reported symptoms and were tested for viral infection. Of those tested, 149 had influenza A virus, 57 had influenza B, and 21 had RSV.
- **How was the data collected?** The activity and sleep data were collected passively from personal Fitbit devices through the Fitbit API. Daily surveys were completed by participants through Evidation's study platform. Test kits were self-administered and then sent for analysis in an approved lab. All collection (passive and active) was consented to by participants through the informed e-consent at the beginning of the study.
- **Over what time-frame was the data collected?** Data were collected over a five month period from December 2020 to June 2020.
- **Does the dataset contain all possible instances?** No, this dataset only contain a subset of the population and possible health outcomes.
- **Is there information missing from the dataset and why?** While key completion metrics were high across the board (see C.8), there is information missing, mostly due to participants not filling in surveys or not wearing their devices. A small number of missing data can be attributed to rejected test kits and invalid lab results.
- **Other comments about data collection?** N/A

- **How is the dataset distributed? Who is supporting/ hosting/ maintaining the dataset?** The dataset is available to qualified researchers on the Synapse platform at .
- **If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?** Yes, all participants were explicitly told what data would be collected and e-consented to participate in the study. The study was approved by Western Institutional Review Board, Inc. (Puyallup, WA).
- **Does the dataset contain information that might be considered sensitive or confidential?** The full dataset includes information about demographics and medical history. We purposefully limited the type of data that was collected to reduce the collection of sensitive data. The dataset is coded, which means that all PII (personal identifiable information) is removed or replaced with random identifiers.

D.2. Hyper-parameter search space

All models in this paper were trained with a randomized hyperparameter sweep using a withheld validation set. For CNN modules, we experimented with kernel sizes as large as 63, stride sizes as large as 256, depths as deep as eight layers, and as many as 32 output channels. For transformer modules, we experimented pre-computed and fixed positional embeddings, up to twelve layers of stacked transformers, and up to nine-head attention. We also tried dropout rates between 0.0 and 0.5. In total, over five hundred model configurations were tested before setting on the final configuration of kernel sizes of 5,5,2, stride sizes of 5,3,2, output channels of 8,16,32, two transformer layers each with four heads, and dropout of 0.4. We tried Adam learning rates from 1 to $1e-6$, and found that $5e-4$ worked best. This relatively small learning rate seemed to be important for limiting overfitting. We also conducted a hyperparameter sweep for XGBoost models, and that $\eta = 1$ and a maximum depth of six worked best. Further, we experimented with window sizes ranging from three to ten days, and found that the model overfitted on both ends of this range, with best performance at seven days. ResNet hyperparameters were borrowed from Ruiz et al. (2021). Our transformer classifier used a learning rate of $2e-3$ and nine blocks. All neural models were trained with early stopping such that training ended once ROC AUC on the validation set decreased for two consecutive epochs.

D.3. Software and Hardware specifications

The full conda environment for these experiments is specified in our github repository: Need to update
All models were trained on a virtual machine with one Nvidia RTX9000, 5 cpu cores, and 64GB of RAM.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All data, including raw data, daily surveys, lab test results, and processed data are available through Synapse at https://www.synapse.org/#!Synapse:syn22803188. The code is available in the project's GitHub repository: https://github.com/behavioral-data/Homekit2020.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This is detailed in the Appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] However, we quantify statistical uncertainty with respect to model rankings in our critical difference plots (Section 5)

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is detailed in the Appendix.

3. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] This is detailed in the Appendix.

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All data, including raw data, daily surveys, lab test results, and processed data are available through Synapse at https://www.synapse.org/#!Synapse:syn22803188. The code is available in the project's GitHub repository: https://github.com/behavioral-data/Homekit2020.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] These are detailed in the Section C of the Appendix.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] This is done in section D of the Appendix (Datasheet for Homekit2020).