Self-Supervised Pretraining and Transfer Learning Enable Flu and COVID-19 Predictions in Small Mobile Sensing Datasets

Mike A. Merrill MIKEAM@CS.WASHINGTON.EDU
Tim Althoff ALTHOFF@CS.WASHINGTON.EDU
University of Washington

Abstract

Detailed mobile sensing data from phones and fitness trackers offer an opportunity to quantify previously unmeasurable behavioral changes to improve individual health and accelerate responses to emerging diseases. Unlike in natural language processing and computer vision, deep learning has yet to broadly impact this domain, in which the majority of research and clinical applications still rely on manually defined features or even forgo predictive modeling altogether due to insufficient accuracy. This is due to unique challenges in the behavioral health domain, including very small datasets ($\sim 10^1$ participants), which frequently contain missing data, consist of long time series with critical long-range dependencies (length$< 10^4$), and extreme class imbalances ($> 10^3$:1).

Here, we describe a neural architecture for multivariate time series classification designed to address these unique domain challenges. Our proposed behavioral representation learning approach combines novel tasks for self-supervised pretraining and transfer learning to address data scarcity, and captures long-range dependencies across long-history time series through transformer self-attention following convolutional neural network-based dimensionality reduction. We propose an evaluation framework aimed at reflecting expected real-world performance in plausible deployment scenarios. Concretely, we demonstrate (1) performance improvements over baselines of up to 0.15 ROC AUC across five influenza-related prediction tasks, (2) transfer learning-induced performance improvements including a 16% relative increase in PR AUC in small data scenarios, and (3) the potential of transfer learning in novel disease scenarios through an exploratory case study of zero-shot COVID-19 prediction in an independent data set. Finally, we discuss potential implications for medical surveillance testing.

Data and Code Availability This paper uses data from the Homekit2020 Flu Study. Of the 5195 participants in this study, 5034 consented to data sharing. This subset is available via Synapse (link withheld to protect anonymity), and is documented in a parallel submission to CHIL along with a benchmark evaluation. We make all code used in this paper 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

Mobile sensing data from phones, watches, and fitness trackers offer an unparalleled opportunity to track complex behavioral changes and symptoms, detect high risk individuals in large populations, and deploy targeted interventions. Because many conditions manifest themselves through behavioral and physiological changes (e.g., reduced activity, disrupted sleep, increased heart rate), leveraging these data could minimize the impact of emerging diseases. Currently, such conditions exact a massive toll (e.g., (Mezlini et al., 2021)), with contagious respiratory illnesses such as COVID-19 (or influenza/flu) rising to the second leading cause of death in the U.S. in January 2022.

Despite the enormous potential and availability of these data for well over a decade, broad and tangible impacts on population health have yet to be realized. For example, consider their limited impact on COVID-19, which reduced gross global product by $28 trillion (International Monetary Fund, 2020); except for contact tracing apps, which do not require predictive modeling, the global COVID-19 response made no significant use of these data beyond research studies.

While neural representation learning approaches have provided transformative performance improvements across Natural Language Processing (NLP) and Computer Vision (CV) (Mikolov et al., 2013; Devlin et al., 2019; Lewis et al., 2019; Lan et al., 2019; Liu et al., 2019; He et al., 2016; Krizhevsky et al., 2012), currently these techniques are rarely adopted for mobile sensing research and applications. In contrast to typical NLP and CV benchmark datasets, mobile sensing data are usually very limited in size (often less than 20 individuals due to arduous and expensive data collection (Xu et al., 2021)), frequently contain missing data (93% of days in our data; e.g. when device is not used or charging), consist of long history time series (>10,000 min-by-min time steps for one week of data) with relevant long range dependencies (e.g. changes in heart rate across multiple days), and feature extreme class imbalances (up to 2760:1 in our data; because most people are not sick on most days) (Xu et al., 2021). Therefore, researchers have often been limited to using small datasets and less data hungry non-neural models such as boosted tree models with hand-crafted features which typically perform worse (e.g., (Xu et al., 2021; Laport-López et al., 2020; 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 this paper, we describe a neural architecture for multivariate time series classification specifically designed for these unique domain challenges. Specifically, (1) this model learns directly from raw minute-level sensor data (in contrast to prevailing use of manually defined features; Section 3.1), (2) leverages novel self-supervised pretraining tasks (Section 3.2) and transfer learning to improve performance in datasets of limited size without requiring additional supervision, (3) directly models potentially informative missingness patterns instead of excluding participants with missing data, (4) captures long-range dependencies across long-history time series through transformer layers (Vaswani et al., 2017), while (5) reducing the input sequence length to these transformer layers through hierarchical feature extraction of convolutional neural networks (CNNs).[1]

Next, we present a framework of best practices for evaluating mobile sensing models (Section 5), which describes (1) how to avoid massively overestimating model performance relative to expected real-world performance, and (2) an approach to reduces the statistical uncertainty introduced by inherent extreme class imbalances (up to 1:2,760 in our evaluation data; Section 4) that is based on jointly comparing model performance across multiple prediction tasks. We then apply this framework to the evaluation of the proposed model across four experiments.

In Experiment 1 (Section 6.1) we evaluate model performance across five single domain prediction tasks related to predicting the flu with FitBit wearable data, and show that CNN encoders, transformer blocks, modeling missingness, and self-supervised pretraining significantly increase predictive performance, up to 0.15 ROC AUC relative to common baselines.

In Experiment 2 (Section 6.2) we compare three novel self-supervised pretraining tasks and show that a task which incorporates basic domain expertise performs best.

In Experiment 3 (Section 6.3) we demonstrate transfer learning of pretrained behavioral representations. Specifically, we simulate 20 separate small data studies with only ten participants each for training. We show that finetuning a pretrained model (trained on a separate self-supervision task on an independent set of participants) on these ten participants outperforms training from scratch with a $\sim$ 16% improvement in precision-recall AUC.

In Experiment 4 (Section 6.4) we extend the previous transfer learning setting to an exploratory case study of zero-shot COVID-19 prediction in a small third party dataset. In this zero-shot paradigm, without any training on COVID-19 cases, the proposed pretrained model achieves a 0.62 ROC AUC, while an XGBoost baseline cannot exceed near-random performance (0.51 ROC AUC).[2] This demonstrates that the pretraining of the proposed model architecture is able to learn generalizable features that enable significant performance improvements across multiple domains (flu and COVID-19).

Finally, we reflect on these advances and potential implications in the the context of the medical literature on surveillance testing (Section 7). Advances in model performance especially on small datasets and novel disease scenarios could support a more

1. Because the full self-attention mechanism of transformers has computational and memory requirements that are quadratic with the input sequence length (Beltagy et al., 2020)

2. Note that statistical power is limited due to the small dataset size, and therefore we first include the repeated simulation study of Experiment 3 to demonstrate robustness of transfer learning performance.

widespread use of mobile sensing data as well as enable rapid deployments in emerging disease scenarios. For example, in the crucial early days of the COVID-19 pandemic, laboratory testing was not widely available, and many positive cases remained undetected. In this setting, a generalizable pretrained model could be fine-tuned on the few test cases already available, and used to identify members of a population who may be infected and should be targeted for additional testing or interventions (Brook et al., 2021; Nestor et al., 2021; Quer et al., 2020). It is promising to note that the proposed model enables predictive performance comparable to some flu and COVID-19 rapid antigen tests (ca. 0.68 to 0.88 ROC AUC (Bachman et al., 2021; Chu et al., 2012)). Still, we emphasize that additional research and validation experiments are needed to support the use of predictive models such as ours in public health strategy and policy.

In summary, our contributions include:

- A neural architecture for multivariate time series classification in mobile sensing and novel set of pretraining tasks (Section 3)
- A framework for evaluating mobile sensing models, which provides best-practices for selecting realistic prediction tasks and mitigating inherent statistical uncertainty during model selection (Section 5)
- An empirical evaluation demonstrating that the proposed approach significantly improves prediction performance on small datasets through transfer learning in both flu predictions and a case study of a novel zero-shot COVID-19 prediction task (Section 6).

We make our model publicly available for use by researchers and practitioners at REDACTED, so that they may use it as an initialization for their own prediction tasks.

2. Related Work

Our model builds upon prior work in neural methods and transfer learning for behavioral sensing and modeling. Our model is the first to learn generalizable feature representations from long-history multivariate time series to enable transfer learning in small datasets.

2.1. Neural Models for Time Series Classification in Mobile Sensing

Behavioral data has been modeled and mined using deep learning techniques across a variety of domains, including human activity recognition (using CNN) (Yao et al., 2017), personalized fitness recommendation (using stacked LSTM (Hochreiter and Schmidhuber, 1997)) (Ni et al., 2019), mood prediction (using RNN, GRU, or autoencoder) (Suhara et al., 2017; Cao et al., 2017; Spathis et al., 2019), stress prediction (using LSTM and autoencoder) (Li and Sano, 2020), health status prediction (using CNN and cross-attention) (Hallgrímsson et al., 2018), and personality prediction (Wu et al., 2020). Two studies experimented with multi-head attention and convolution as we do here, but neither paper applies this architecture to transfer learning (Song et al., 2018; Tang et al., 2021). Liu et al. (2022) apply a CNN autoencoder to raw sensor data to predict COVID-19, but do not experiment with transformer layers nor transfer learning as we do here.

Until very recently, the state of the art in time series classification has eschewed deep learning in favor of more traditional statistical learning methods (Fawaz et al., 2019). Fawaz et al. (2019) propose a set of benchmark datasets and tasks for time series classification, but relative to the minute-level time series we model here these datasets are shorter (at most 2,000 observations in length), do not contain missing data, and are mostly univariate.

2.2. Self-Supervised Learning in Behavioral Modeling

In the broader field of self-supervision for time series classification, the most relevant work is Zerveas et al. (2021), who use a simple linear projection to shrink the input multivariate time series to the scale supported by transformers. Zhang et al. (2019) use an LSTM to learn self-supervised representations of multivariate timeseries, but apply this method to anomaly detection. Transfer learning remains a "grand challenge" for mobile sensing (Wang et al., 2019), and has been explored in human activity recognition (Ma et al., 2020), stress and mood prediction (Jaques et al., 2017; Li and Sano, 2020), and forecasting adverse surgical outcomes in an ICU (Chen et al., 2020). Hallgrímsson et al. (2018) use a CNN autoencoder to forecast heart rate from steps and sleep data, but do not predict acute events like viral infection. Kolbeinsson et al. (2021) pretrain

transformers to auto-regressively predict one of several handcrafted features on a day given the previous day's aggregated sensor data, but do not model minute-level raw time series to predict a multitask array of features as we do here. Furthermore, none of these applications focus explicitly on model transfer to small datasets. Tang et al. (2021) study performance on small datasets, but unlike our work do not evaluate with a zero-shot, out-of-domain task.

3. Methods

Here, we detail our model which is composed of a CNN encoder for learning hierarchical and temporal features, effectively reducing the dimensionality, and a transformer for learning potentially long-range relationships between these features. Then, we motivate and describe a set of self-supervised pretraining tasks that can be used to boost overall model performance. [3]

3.1. Our Model

Our Model is composed of a convolutional encoder, a stack of transformer blocks, and a final densely connected linear layer that is used for classification. Intuitively, the convolutional encoder learns a compressed, hierarchical feature representation of the raw time series data, while the transformer learns relationships between these features. An overview of our model architecture is given in Figure 1.

Notation. Formally, we define a given input sensor stream as $x_i \in \mathbb{R}^{m\times 1}$, where m is the length of the time series, and x_{it} is the value of sensor stream i at time t. We assemble $X = (x_0, .., x_n) \in \mathbb{R}^{m\times n}$ as a multivariate time series of n streams in a given user's data.

Convolutional Encoder. The convolutional encoder learns a temporal, hierarchical feature representation of the raw sensor data. Given the input multivariate time series X, we stack q convolutional layers. In the simplest case, when stride and kernel size are not considered, the output of the j^{th} channel C_j with input size $(C_{\text{in}}, L_{\text{in}})$ is:

$$\text{out}\,(C_j) = \text{bias}\,(C_j) + \sum_{k=0}^{C_{in}-1} \text{weight}\,(C_j, k) \star \text{input}\,(k)$$

where $\star$ is the cross-correlation operator, and weight and bias reflect learned parameters unique to each channel and each layer. Between layers we apply ReLU and batch-norm, which limit overfitting. We denote the final output of the CNN Encoder as $\overline{X}$, which has dimensionality $(C_{\text{out},q}, L_{\text{out},q})$.

Transformer Blocks. Intuitively, this module learns relationships between the features produced in the final output of the CNN encoder. Our model uses a stack of u transformer blocks, each composed of r attention heads and a feed forward layer. We take the output of the final layer, E, to be the learned representation of the input time series:

$$\begin{aligned} h_0 &= \overline{X}^T + W_p \\ h_i &= \text{TransformerBlock}(h_{i-1}), i \in 1...u \\ h_i^{norm} &= \text{LayerNorm}(h_i) \\ E &= h_u^{norm} \end{aligned}$$

Where W_p is a learned positional embedding matrix.

Allowing for Missing Data. Researchers frequently report missingness as an obstacle to adopting deep learning techniques. In our dataset, 93% of days contain at least a minute of missing data. Accordingly, we model missingness by replacing missing values with zeros and including a binary flag for each of the sensor streams which encodes if the sensor reading is missing in that timestep.

Training. We train our model with the Adam optimizer (Kingma and Ba, 2017) and cross entropy loss. Details about hyperparameter tuning are available on the project's github.

Note on explicitly modeling class imbalance. In writing this paper we experimented with several common techniques for modeling imbalanced classification problems, including focal loss (Lin et al.) and balanced cross entropy loss. We note that these methods did not significantly improve performance, perhaps because the difficulty of the underlying classification problem dwarfs the difficulty imposed by class imbalance. Nonetheless, as we show in Section 5.2, it is important to evaluate model performance on these tasks with the imbalance in mind.

3. We would like to make it abundantly clear that this paper is not the first to combine CNNs with transformers or apply self-supervised learning to time series data. Instead, the core architectural contribution of this paper is the application of these techniques in a unified system to a new domain to address the challenges of modeling behavioral time series data.

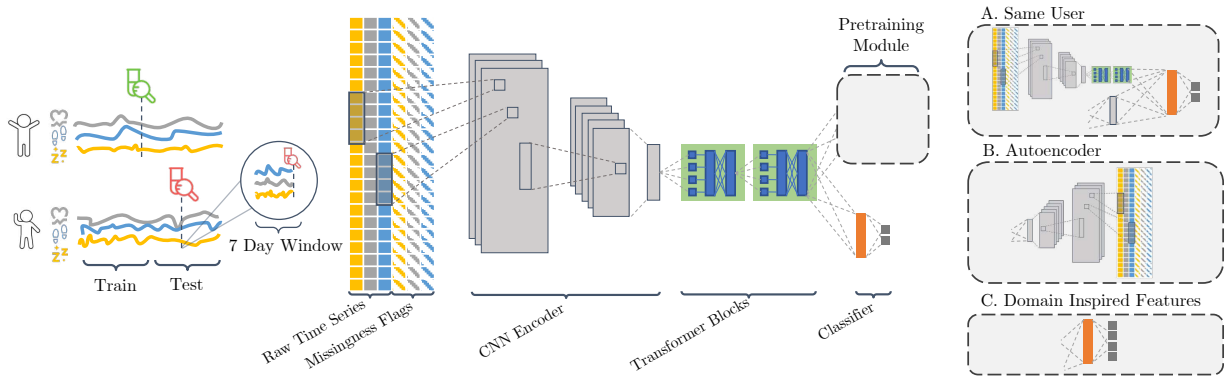


Figure 1: Our Model (Section 3.1) combines a CNN encoder for learning hierarchical and temporal features from raw time series data and a transformer for learning long-range relationships between these features. Additionally, we provide three novel self-supervised pretraining tasks for learning from unlabeled data (Section 3.2).

3.2. Self-Supervised Pretraining Tasks

Much like in computer vision and NLP, labeled behavioral health data are expensive to collect at scale because labels require costly testing infrastructure. Transfer learning through self-supervised pretraining helps models learn generalizable representations from unlabeled data, where the status quo is only being able to learn from limited labled data. Here, we propose three techniques for self-supervised pretraining for behavioral data.

Same User. Prior work indicates that data from the same user on different days are often highly correlated relative to data from other users (Wang et al., 2016). Drawing inspiration from next sentence prediction tasks in NLP (Logeswaran and Lee, 2018), we hypothesize that a model that is trained to encode the differences between users may learn useful representations of behavioral data. To this end, we construct a dataset of one million pairs of (non-overlapping) windows from the same user, and one million pairs of windows from different users. We use the same encoder to generate embeddings for each of the windows in the pair, concatenate the embeddings, and use a linear layer to classify whether the pair of windows were from the same user (Figure 1A).

Autoencoder. For this pretraining task, we add a CNN decoder to the end of our model and use a mean-squared error objective to learn a reconstruction of the input time series from our model's lower dimensional embedding (Figure 1B). For simplicity's sake our decoder is a reflection of the encoder, i.e. it has the same architecture but with its one dimensional convolutions replaced with one dimensional deconvolutions and with a decreasing number of channels such that the final output has the same dimensionality as the original input.

Domain Inspired Features. As previously mentioned, the majority of prior work in behavioral modeling has focused on classification tasks with handcrafted features. While neural minute-level models may achieve superior performance than simple classifiers trained on these features, there is nonetheless a large body of work supporting the utility of handcrafted features in sensing (Xu et al., 2021; Laport-López et al., 2020; 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). For this pretraining task, we ask the model to perform a multiple regression to predict the daily features in Table A1 on the final day of the seven day window (Figure 1C). Intuitively, there may be other, less obvious yet highly informative orthogonal features that our model could learn in order to reconstruct these higher level features. This task also has the added benefit of allowing us to inject expertise into the model. Since these features are calculated from the raw data (and in fact are mostly available through the Fitbit API) and do not require any exogenous labels this task is fully self-supervised.

Feature	Description
Number of participants	5196
Average number of days of data	114
Mean % of missing data per day	9.8%
Mean age (±SD)	37.7 (10.2)
Portion Female	72%
Mean BMI (±SD)	30.3 (20.3)
Number of US States Represented	50

Table 1: Summary statistics for the Homekit Flu Monitoring Study (Section 4)

4. Dataset

Our dataset consists of 591k user-days of Fitbit data collected from 5196 participants in the Homekit Flu Monitoring Study over the course of six months. Each minute the devices recorded the participant's total steps, average heart rate, and binary flags indicating if the participant was sleeping, awake, or in bed. Participants also completed daily surveys which asked if they were experiencing flu symptoms, including coughing, chills, fever, and fatigue. When a participant indicated that they were experiencing a cough and one other symptom, they were asked to self-administer a nasal swab test kit, which was then mailed to a lab for PCR analysis. Table 1 contains summary statistics for this study.

5. Challenges in Evaluation

There are few, if any, established best practices for evaluating behavioral models (Nestor et al., 2021; McDermott et al., 2021). Given this lack of guidance, current evaluation paradigms vary significantly across studies. Here, we identify two common challenges to evaluating behavioral models in health and propose accompanying solutions, which we use to compare models in Section 6. In summary, evaluations for behavioral models in healthcare should:

- Replicate genuine conditions, such as only using data from the past to inform predictions, and be tolerant to endemic missing data (Section 5.1).
- Faithfully quantify statistical significance, in particular when condition positive examples are rare, as is often the case in diagnostic testing (Section 5.2).

5.1. Problem: Evaluations in artificial settings may lead to misleading performance estimates

Without a clear health application in mind from the outset it can be difficult for researchers to define tasks which faithfully replicate "real world" conditions. It is not uncommon for models to:

- train on data from the future, e.g. by using data from one user at the end of the data collection period to inform predictions about another user at the beginning (Wang et al., 2016),
- use data collected in laboratory settings with limited ecological validity (Ismail et al., 2020),
- make predictions only if a user supplies sufficient data by using a device frequently (Malik et al., 2020; 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 with endemic missing data (Nestor et al., 2021; Ismail et al., 2020).

Solution: Situate tasks around plausible healthcare scenarios. Here, 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?

Such a scenario arises in surveillance testing, where a population is frequently tested and positive individuals are asked to undertake additional testing or self isolate (Mercer and Salit, 2021). 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 users 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).

5.2. Problem: Predicting rare events limits statistical power and makes model selection inherently challenging

In mobile sensing for public health, relevant events are often fairly rare as intuitively most people are not sick most days. For example, one useful application of surveillance testing for respiratory viral infections is that if an individual tests positive they can self isolate and limit the spread of the infection to others. The CDC estimates that the average American has a 10% chance of a symptomatic flu infection in a 365 day period, implying that the probability of an American receiving an initial positive flu diagnosis on a given day is roughly 0.027% (cdc, 2021a). This corresponds to a 1:3,703 class imbalance, similar to the 1:2,760 in our evaluation dataset.

Modeling challenges aside, these extreme class imbalances make comparing model performance difficult as they limit statistical power and lead to large confidence intervals across many common test statistics. For example, for the Delong Test, a common test for comparing the ROC AUCs of two classfiers, 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). Intuitively this variance is minimized, and statistical power maximized, when $m = N/2$ (a 1:1 class balance), and variance is maximized when $m = 1$.

Empirically, uncertainty can be quite high on realistic tasks, with peer studies of COVID-19 and flu detection reporting confidence intervals as high as ± 0.1 ROC AUC (Quer et al., 2020). Such statistical uncertainty makes it difficult to compare models, as extreme improvements in predictive performance on individual tasks are required to make strong claims about methodological progress.

As outlined in Section 5.1, many studies of interesting phenomena such as COVID-19 massively subsample true negatives to artificially deflate this class imbalance (Quer et al., 2020). This creates a much simpler (but unrealistic) task, since higher false positive rates do not massively impact overall performance (Haibo He and Garcia, 2009).

Solution: Aggregate performance across multiple tasks to increase statistical power. Here, rather than directly compare the performance of models on *individual* tasks, we instead jointly compare the relative performance of models across *all* tasks to improve statistical power. Intuitively, if a model performs best on all tasks, but not with high statistical significance on any one test, the probability that the model performance is indeed the same as all others is low. Specifically, we employ a Critical Difference plot (Brazdil and Soares, 2000), which first uses Friedman's statistic (Friedman, 1940) to test the null hypothesis that there is no difference between the relative performance of models, and then deploys pairwise significance tests (e.g. Wilcoxon signed rank) between classifiers. This method, used here for the first time in mobile sensing for epidemiology, allows us to make statistically sound claims about our model's improvement over other common techniques without simplification of the underlying tasks (Section 6).

6. Empirical Evaluation

Here, we define five realistic prediction tasks and compare Our Model's performance against three representative baselines inspired by prior work (Section 6.1). In Experiment 1 we evaluate the performance between tasks through the framework defined in Section 5 to show that Our Model outperforms these baselines. Next, Experiment 2 compares pretraining methods for behavioral data to show that a method which integrates simple domain knowledge performs best (Section 6.2). Experiment 3 in Section 6.3 then shows that in simulated settings with limited training data, pretraining provides an average 0.04 ROC AUC performance boost relative to a non-pretrained model. Finally, we use a small, independently collected Fitbit dataset to illustrate that features learned by our model on flu prediction generalize to COVID-19 prediction in a zero-shot task in Experiment 4.

6.1. Experiment 1: Realistic Single Domain Prediction Tasks

We evaluate methods on five behavioral modeling tasks. Below "severe" constitutes a three or more on a four point Likert scale.

- **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 and asked to self-isolate in the event of a positive result (Chu et al., 2020; Fusco et al., 2020).

	Flu Positivity		Severe Fever		Severe Cough		Severe Fatigue		Flu Symptoms	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
XGBoost (Day Level)	0.708	0.003	0.741	0.013	0.704	0.018	0.708	0.032	0.647	0.044
LSTM	0.674	0.001	0.733	0.006	0.649	0.008	0.710	0.017	0.606	0.026
ResNet	0.551	0.001	0.701	0.004	0.629	0.007	0.686	0.014	0.629	0.036
CNN	0.860	0.002	0.801	0.015	0.690	0.008	0.699	0.016	0.612	0.024
CNN-Transformer	0.884	0.007	0.790	0.039	0.697	**0.023**	0.713	0.038	0.640	0.042
CNN-Transformer Pretrained	**0.887**	**0.010**	**0.818**	**0.056**	**0.708**	**0.023**	**0.758***	**0.074***	**0.671***	**0.066***
Class Balance	1:2,760		1:643		1:132		1:78		1:37	

Table 2: Results on all tasks for our model. *Indicates $p < 0.05$ (Delong). Note that while substantial class imbalance precludes statistically significant results on some tasks ("Flu Positivity", "Severe Fever", and "Severe Cough"), Our Model performs better than all baselines and ablations when jointly evaluating performance across *all* tasks to increase statistical power (Figure 2).

- **Severe Fever:** Will the participant report a severe fever today?
- **Severe Cough:** Will the participant report a severe cough today?
- **Severe Fatigue:** Will the participant report severe fatigue 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.

For these tasks, we follow our aforementioned evaluation best practices (Section 5.1) by training with data before the midpoint of the flu season (February 10th, in our case), and testing and evaluating models on data after the midpoint (as only data for one flu season is available). Furthermore, we make a prediction for every user on every day regardless of data quality, including predictions for users with no true positive labels.

In each case, we compare our model to the following baselines:

- **XGBoost:** How well does our model perform relative to a non-neural baseline? Boosted decision trees are frequently used in many sensing studies because they are supported by common, easy to use libraries and often achieve strong performance out-of-the-box (Xu et al., 2021). Since boosted trees expectedly do not scale well to the thousands of observations in our raw time series data, we compute a set of commonly used features for each day in the window, and then concatenate these features for a final input. 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)). A list of all features is available in Table A1.
- **LSTM:** How well does a recurrent model perform on this task? LSTMs are strong baselines in time series classification Ruiz et al. (2021) and EEG processing Craik et al. (2019).
- **ResNet:** How well does a competitive neural model for time series classification perform on our task? While ResNet typically underperforms the state of the art in most computer vision tasks, 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).
- **CNN:** How important are the transformer layers to our model's performance? To answer this question, we removed the transformer blocks from our model and passed the CNN's final output directly to a linear layer. 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).

- **CNN-Transformer:** How important are pretraining and missingness flags to our model's performance? For this ablated model, we pass the CNN's final output to a transformer, but do not apply any pretraining method and do not include missingness flags.

We do not include a "transformer only" baseline (i.e., our model without the CNN encoder) because multi-head attention scales quadratically with the input length, making it computationally infeasible to perform such an experiment on a multi-day timeseries window (i.e., minute level data on a seven day window produces a 10,080 dimensional vector), which exceeds common context sizes in transformer models on commodity GPUs (Beltagy et al., 2020).

Results. Our model outperforms all baselines on every task (Table 2), which indicates that our method is a meaningful improvement over state of the art classifiers for behavioral data. Here we focus on precision-recall AUC, since the metric is typically more informative in cases of extreme class imbalance (Saito and Rehmsmeier, 2015). Through Delong's test we find significant improvements in ROC AUC and PR AUC on the "Severe Fatigue" and "Flu Symptoms" tasks at $\alpha = 0.05$. As outlined in Section 5, we employ Friedman's test and pair-wise Wilcoxon signed-rank tests to compare performance across tasks, and find that Our Model significantly outranks XGBoost, CNNs, and CNN-Transformers at the best-practice parameter $\alpha = 0.1$ (Brazdil and Soares, 2000), as it ranks first across all tasks. A critical difference plot is available in Figure 2, which shows that Our Model is the best performing model overall, and that there is no statistically significant difference in the rankings of XGBoost, CNN, and the (non-pretrained) CNN-Transformer. We also experiment with the model's performance at modest levels of missing data, and find that it compares favorably to XGBoost (e.g. over 0.9 ROC AUC on the "Flu Positivity" task even with 20%-30% of data missing; Figure A.1). A complete summary of results is available in Table 2.

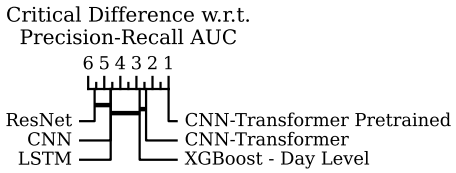


Figure 2: Critical Difference Plot (Brazdil and Soares, 2000) between models at $\alpha = 0.1$. Numbers indicate each model's average ranking on the single domain prediction tasks (Section 6.1), while the thick dark line connects models which are not significantly different from one another. This demonstrates that Our Model, which uses pretraining and models missing data, significantly outperforms ResNet, CNNs, XGBoost, and CNN-Transformers across tasks (average rank=1.0).

6.2. Experiment 2: Comparison of Self-Supervised Pretraining Methods

Next we compare the three pretraining techniques proposed in Section 3.2 on the "Flu Symptoms" task (Section 6.1). This "Flu Symptoms" task has the least extreme class imbalance (1:37) and therefore yields highest statistical power to differentiate model performance. We use the following pretraining method:

1. Pretrain the model using all seven day windows in the train dataset.
2. Freeze the model's CNN and transformer layers. If the pretraining technique used a classification head, randomize its parameters. If instead a regression head was used, replace it with a randomly initialized classification head.
3. Finetune the model on the target task ("Flu Symptoms", in this case).

For all experiments, we use all of the model features described in Section (3.1) (i.e., the convolutional encoder, transformer blocks, and missingness flags). For comparison, we include a "No Pretraining" baseline, which shows the performance of a randomly initialized model.

Results. ROC and Precision Recall curves for this experiment are available in Figure 3. "Domain Inspired Features" pretraining, which trains the model to predict a pre-computed set of handcrafted features (Section 3.2), significantly outperforms other pretraining techniques *and* a randomly initialized model with a 16% improvement in PR AUC. Notably, the model pretrained on the "Same User" task does significantly worse than the others. One plausible explanation is that by learning to embed windows from the same user in the same region of the latent space, the

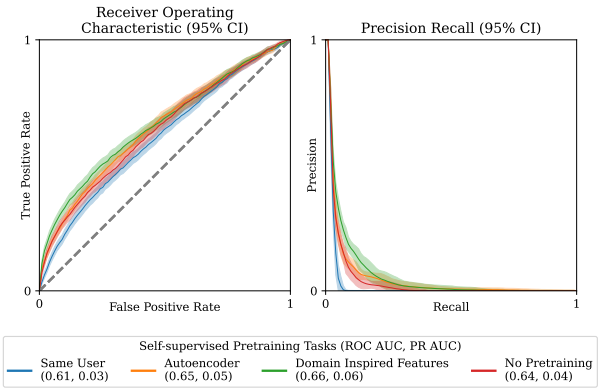


Figure 3: Comparison of self-supervised pretraining tasks (Section 3.2) on the "Flu Symptoms" task. The "Domain Inspired Features" task, which integrates domain knowledge, performs best.

model sacrifices its ability to distinguish "unusual" (e.g. flu positive) windows for a given user, since these windows would ordinarily be much further away in the latent space.

We additionally compared this pretraining-fine tuning approach to a multitask learning where both the pretraining and the target prediction objective are optimized *concurrently*. We repeated this comparison for each of the pretraining tasks (Section 3.2) in combination with the "Flu Symptoms" task. This strategy produced no meaningful improvement over the randomly initialized model, indicating that unsupervised pretraining is a superior paradigm for this setting. In addition, the pretraining-finetuning paradigm enables us to separate these two steps across two datasets, especially when the target dataset is relatively small. This is the focus of the next two experiments.

6.3. Experiment 3: **Transfer learning improves flu prediction performance on small datasets in a repeated simulation study**

Labeled behavioral data is often prohibitively expensive to collect, particularly in the context of public health where ground-truth labels require costly testing infrastructure and study management. Accordingly, many studies from prior work operate on data with on the order of dozen participants (Xu et al., 2021). In this regard, one promising application of generalizable self-supervised pretraining is that models could leverage large unlabeled datasets to improve predictive power in settings with limited labeled training data. Here, we repeatedly simulate such settings to robustly investigate whether such transfer learning leads to performance improvements.

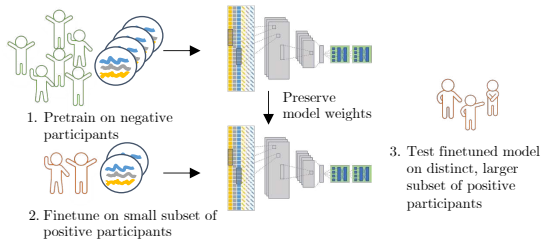
First, we isolate all 4,989 study participants who never tested positive for the flu. We treat this set as a large, unlabeled dataset which we use to pretrain our model on the self-supervised "Daily Features" task (Section 3.2). We then take the remaining 206 users who *did* test positive at some point during the study, and randomly split this set into twenty folds of ten or eleven users each. This ensures that source and target domain share no participants in common. We provide an overview of this split in Figure 4(*a*).

Next, for each fold we finetune the model on the supervised "Flu Positivity" task using data from the fold, and evaluate it on the users in the remaining nineteen folds. We choose this task as it mirrors the zero shot setting in the external dataset of Experiment 4. In both of these settings, all test subjects tested positive at some point and the predictive model attempts to predict on which day they do so. This process simulates finetuning the model with fewer than a dozen users' data. We compare this approach to two non-pretrained models that only have access to the smaller target domain dataset: CNN-Transformer, and XGBoost trained on manually defined features (Table A1).

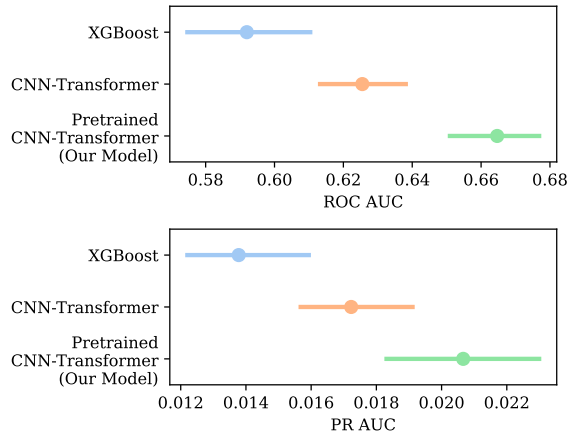
Results. We find that our pretrained model outperforms non-pretrained models on the "Flu Positivity" task when trained on fewer than a dozen participants (Figure 4(*b*)). Pretraining alone increases average performance from 0.626 ROC AUC to 0.665, and 0.017 PR AUC to 0.021 (both $p < 0.05$, Mann-Whitney U). This indicates that Our Model can learn generalizable features from unlabeled data.

6.4. Experiment 4: **Zero-shot COVID-19 prediction in a small external dataset**

It is plausible that a self-supervised pretrained model, which in Experiment 1-3 showed good performance on flu related tasks, could support non-random predictive performance in a zero shot setting for COVID-19? Both diseases are respiratory viral infections and and may trigger similar behavioral and physiological responses (e.g., a change in resting heart rate around symptom onset (Shapiro



(*a*) The cross validation split for our simulation of small datasets (Section 6.3)



(*b*) Performance on the "Flu Positivity" task (Section 6.1) using training data from ten participants. Neural models outperform XGBoost, and pretraining yields additional improvement. Horizontal lines represent 95% confidence intervals (all $p \leq 0.05$; Mann-Whitney U).

Figure 4: Simulating "small data" scenarios using training data from only ten participants (Section 6.3).

	XGBoost	Our Model
Zero-shot PR AUC	0.005	0.018
Zero-shot ROC AUC	0.51	0.68

Table 3: Performance on zero-shot COVID-19 Prediction (Section 6.4). Our model's superior performance shows that CNN-Transformers pretrained on the "Domain Inspired Features" task (Section 3.2) learn generalizable features.

et al., 2021)). We use a small, independently collected dataset of Fitbit recordings and COVID-19 test results to show that Our Model can learn representations which generalize to entirely unseen diseases. This dataset contains 1470 total days of data for 32 individuals who tested positive with COVID-19 (Mishra et al., 2020). The original study uses a retrospective prediction task with no train/test split, and so it is not possible to make a direct comparison between our model and theirs, but this dataset allows us to test performance on an unseen disease.

We pretrain our model with the "Domain Inspired Features" task (Section 3.2) and finetune it on the "Predict Flu Positivity" task (Section 3.2). Note that this is the same configuration as "Our Model" in Table 2. Then, with no additional supervision we use the model to predict COVID-19 positivity in the small, external dataset. As a zero-shot baseline we calculate a set of day-level features (Table A2) from these data "and" our original flu dataset (Section 4) and train XGBoost with these features on the "Flu Prediction" task. Neither Our Model nor the XGBoost baseline is exposed to any data from the COVID-19 dataset during training.

Results. Our Model outperforms XGBoost on this zero-shot task, achieving 0.68 ROC AUC, while XGBoost predicts at 0.51 ROC AUC (random chance) (Table 3). This illustrates the feasibility of pretrained CNN-Transformers for novel disease prediction.

7. Discussion & Conclusion

During the COVID-19 pandemic over-the-counter antigen tests have been effective in surveillance testing, but the frequency of tests, more so than their sensitivity, remains a barrier to success in mitigating spread (Larremore et al., 2020). This paper presents a framework for evaluating mobile sensing methods for frequently predicting respiratory viral infections. While there are limitations to this study, our results show performance on par with COVID-19 rapid diagnostic tests in similar surveillance settings. More research is needed to demonstrate similar performance levels in larger studies. Nonetheless, our findings suggest that mobile sensing predictions can complement rapid antigen testing or trigger additional testing. Our results indicate that pretraining, transformer self-attention, modeling missing data, and transfer learning are effective techniques in learning generalizable behavioral representations for mobile sensing.

Acknowledgments

The authors thank the HomeKit2020 team, including Matthew Thompson and Barry Lutz, for sharing their dataset for research purposes. 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 Audere and the Biomedical Advanced Research and Development Authority (BARDA Contract Number 75A50119C00036). 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

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

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

Christine M. Bachman, Benjamin D. Grant, Caitlin E. Anderson, Luis F. Alonzo, Spencer Garing, Sam A. Byrnes, Rafael Rivera, Stephen Burkot, Alexey Ball, James W. Stafford, Wenbo Wang, Dipayan Banik, Matthew D. Keller, David M. Cate, Kevin P. Nichols, Bernhard H. Weigl, and Puneet Dewan. Clinical validation of an open-access SARS-COV-2 antigen detection lateral flow assay, compared to commercially available assays. *PLOS ONE*, 2021. ISSN 1932-6203.

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

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.

Cara E. Brook, Graham R. Northrup, Alexander J. Ehrenberg, Jennifer A. Doudna, and Mike Boots. Optimizing COVID-19 control with asymptomatic surveillance testing in a university environment. *Epidemics*, 2021.

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.

Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D Leow. Deepmood: modeling mobile phone typing dynamics for mood detection. In *KDD*, 2017.

Hugh Chen, Scott Lundberg, Gabe Erion, Jerry H Kim, and Su-In Lee. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *arXiv:2002.04770*, 2020.

Haitao Chu, Eric T. Lofgren, M. Elizabeth Halloran, Pei F. Kuan, Michael Hudgens, and Stephen R. Cole. Performance of rapid influenza H1N1 diagnostic tests: a meta-analysis. *Influenza and Other Respiratory Viruses*, 2012.

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.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. 16(3):031001, 2019. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2552/ab0ab5. URL https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5.

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 (1)*, 2019.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. 33(4), July 2019. ISSN 1384-5810, 1573-756X. arXiv: 1809.04356.

Milton Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. 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.

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 E.A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 2009. ISSN 1041-4347.

Haraldur T Hallgrímsson, Filip Jankovic, Tim Althoff, and Luca Foschini. Learning individualized cardiovascular responses from large-scale wearable sensors data. *NeurIPS ML4H*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. (arXiv:1512.03385), December 2015. arXiv:1512.03385 [cs] version: 1 type: article.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

International Monetary Fund. A Crisis Like No Other, An Uncertain Recovery, 2020.

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

Natasha Jaques, Ognjen (Oggi) Rudovic, Sara Taylor, Akane Sano, and Rosalind Picard. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

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.

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), March 2022. doi: 10.1128/jcm.02070-21. Publisher: American Society for Microbiology.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Francisco Laport-López, Emilio Serrano, Javier Bajo, and Andrew T. Campbell. A review of mobile sensing systems, applications, and opportunities. *Knowledge and Information Systems*, 2020.

Daniel B. Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M. Burke, James A. Hay, Tambe Milind, Michael J. Mina, and Roy Parker. Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance. *medRxiv*, September 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019.

Boning Li and Akane Sano. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *IMWUT*, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection.

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*, 2020.

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*, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. 2018.

Yuchao Ma, Andrew T Campbell, Diane J Cook, John Lach, Shwetak N Patel, Thomas Ploetz, Majid Sarrafzadeh, Donna Spruijt-Metz, and Hassan Ghasemzadeh. Transfer learning for activity recognition in mobile health. 2020.

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.

Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 2021.

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.

Aziz Mezlini, Allison Shapiro, Eric J Daza, Eamon Caddigan, Ernesto Ramirez, Tim Althoff, and Luca Foschini. Estimating the burden of influenza on daily activity at population scale using commercial wearable sensors. *medRxiv*, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

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*, 2020.

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. *Nature Digital Medicine*, 2020.

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.

Jianmo Ni, Larry Muhlstein, and Julian McAuley. Modeling heart rate and activity data for personalized fitness recommendation. In *WWW*, 2019.

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, Jennifer M. Radin, Matteo Gadaleta, Katie Baca-Motes, Lauren Ariniello, Edward Ramos, Vik Kheterpal, Eric J. Topol, and Steven R. Steinhubl. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine*, 2020.

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), March 2021. ISSN 1573-756X.

Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 2015.

Allison Shapiro, Nicole Marinsek, Ieuan Clay, Benjamin Bradshaw, Ernesto Ramirez, Jae Min, Andrew Trister, Yuedong Wang, Tim Althoff, and Luca Foschini. Characterizing covid-19 and influenza illnesses in the real world via person-generated health data. *Patterns*, 2021.

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.

Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*, 2018.

Dimitris Spathis, Sandra Servia-Rodriguez, Katayoun Farrahi, Cecilia Mascolo, and Jason Rentfrow. Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data. In *KDD*, 2019.

Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *WWW*, 2017.

Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *arXiv:2102.06073*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 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.

Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. Representation learning on variable length and incomplete wearable-sensory time series. 2020.

Xuhai Xu, Jennifer Mankoff, and Anind K. Dey. Understanding practices and needs of researchers in human state modeling by passive mobile sensing, 2021.

Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *WWW*, 2017.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A Transformer-based Framework for Multivariate Time Series Representation Learning. In *KDD*. August 2021.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V. Chawla. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *AAAI*, July 2019.

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. Reproducability Appendix

A.1. Hyperparameter Tuning

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. The LSTM baseline uses three stacked layers with a hidden size of 128. For transformer modules, we experimented with 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 found 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.

Feature	Description
Resting HR	Avg. heart rate (HR) while still
Main Minutes in Bed	Longest span of minutes 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	Indicator for missing HR data
Missing Sleep	Indicator for missing sleep data
Missing Steps	Indicator for missing steps
Missing Day	Indicator for missing all data

Table A1: Summary of manually defined features, 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.

Feature	**Description**
Resting HR 95^{th} Pct	95^{th} percentile of resting HR
Resting HR 50^{th} Pct	50^{th} percentile of resting HR
Resting HR std.	Standard deviation of resting HR
Awake HR 95^{th} Pct	95^{th} percentile of HR while awake
Steps Streak 95^{th} Pct	95^{th} percentile of continuous steps
Steps Streak 50^{th} Pct	50^{th} percentile of continuous steps
Total Minutes in Bed	Number of minutes spent in bed
Sleep Minutes	Number of minutes spent asleep
Total Steps	Total number of steps
Missing HR	Indicator for missing HR data
Missing Sleep	Indicator for missing sleep data
Missing Steps	Indicator for missing steps
Missing Day	Indicator for missing all data

Table A2: Summary of manually defined features used for XGBoost baseline in the zero shot experiment (Section 6.4) 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.

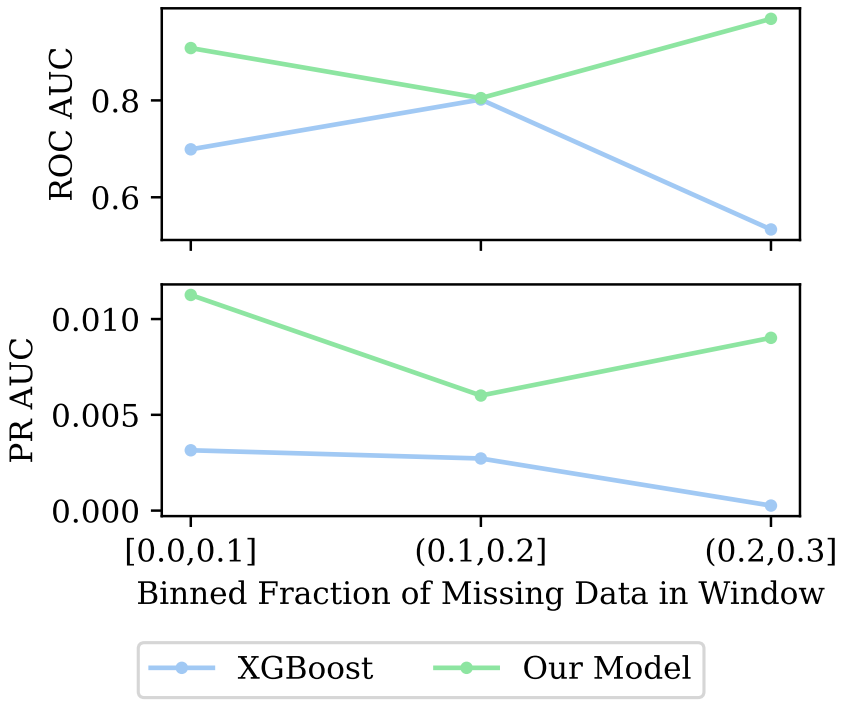


Figure A.1: Performance on the "Flu Positivity" task for binned levels of missing data. Missingness is defined as the fraction of minutes with heart rate data over the duration of the accompanying seven day window. 90% of labels have less that 30% missingness, making positive labels sparse, and so we do not compare methods past this threshold.