Identifying Optimal Case-Onset Points for Early Detection of Influenza-like Illness

Chunting Zheng[1], Edward Ho[1], Shannon Rumsey[1], Jennifer Park[1], Nealson Setiawan[1]
Advisors: Arinbjörn Kolbeinsson[2], Eric Daza[2], Megan Elcheikhali[1]
Sponsors: [1]University of California, Santa Barbara; [2]Evidation Health

evidation

UC SANTA BARBARA | Data Science Initiative

Introduction

- Detecting infectious diseases, such as COVID-19, early can accelerate case isolation and break chains of infection.
- Inconsistent ground-truth labels, i.e., the time-point at which an individual changes labels from healthy to infected, have a significant effect on model training.
- **We aimed to find optimal case-onset for early detection using home testing kit data and fitbit data.**

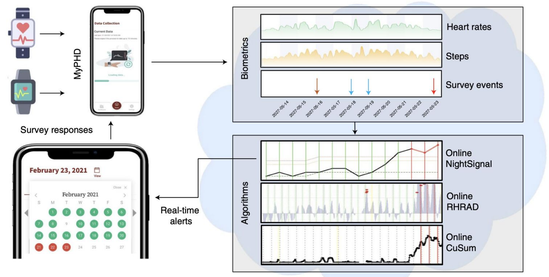

Fig 1. Participants with a Fitbit were asked to share their wearable data. Along with this data, Fitbit provides API to multiple processed features that they have pre-calculated.[1]

Data

- Fitbit data
  - Collected between February and May 2020
  - Contained 5229 individuals who wore wearable (Fitbit) devices
  - Data such as heart-rate, sleep, and exercise were collected[3]
- Survey data
  - Sent to participants daily
  - Recorded if participant was experiencing Influenza-like symptoms
- Lab data
  - Sent self-administered test if participant experienced 2 or more symptoms on the survey
  - Laboratory analyzed PCR test

Multiple Imputation by Chained Equations

47% percent of the data has missing entries, therefore it is imperative that we impute these values and we will do so using MICE.

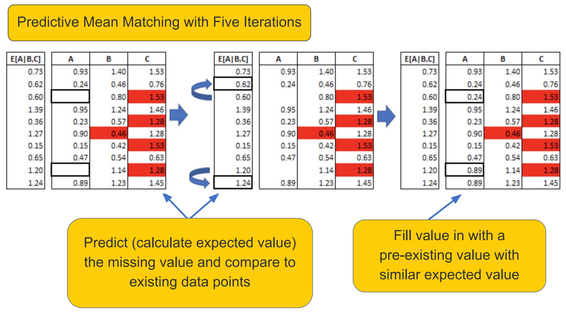

Fig 2. Illustration of the predictive mean matching process that Multiple Imputation by Chained Equations (MICE) uses.[5]

Methodology

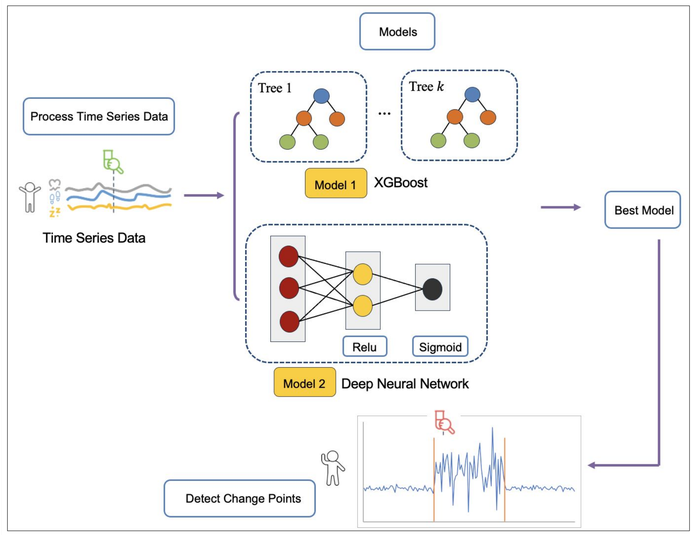

Fig 3. Overview schematic of method. Machine learning model was developed to predict the risk of an individual having a respiratory viral infections, and changepoints were detected.

XGBoost

- **Objective**
  - The boosted tree method to account of dependence between day-to-day observations and changes.
- **Method**
  - Predictors included resting heart rate, total minutes spent in bed, activities calories exerted, time indicators, and aggregated features.
  - A 70%:30% split of data was used for the training and testing datasets.
  - During the model training, observations near the trigger date were exponentially upweighted.

Deep Neural Network

- **Objective**
  - A more complex, robust method to hopefully increase predictive accuracy.
- **Method**
  - One hidden layers utilizing a Relu activation function and a final output layer using the sigmoid function. After each layer, a 0.1 dropout was added to curb overfitting.
  - Model performed best when only containing three predictors: resting heart rate, total minutes spent in bed, and activities calories exerted.

Model Selection and Changepoint Detection

- The performance of the ML models were measured by the area under the receiver operating characteristic curve (AUROC) in the training and testing datasets.
- We chose the best performing ML model that has the higher ROC-AUC score.
- We applied online changepoint detection on the risk of an individual having a respiratory viral infections predicted by the better ML model.
- The detected changepoints will likely be the onset points.

Results

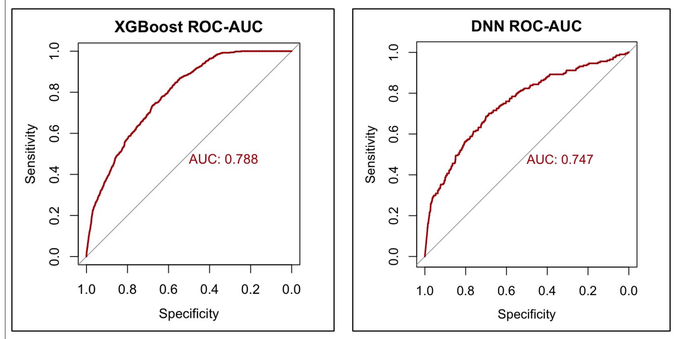

Fig 4. ROC-AUC score of XGBoost (A) and DNN (B).

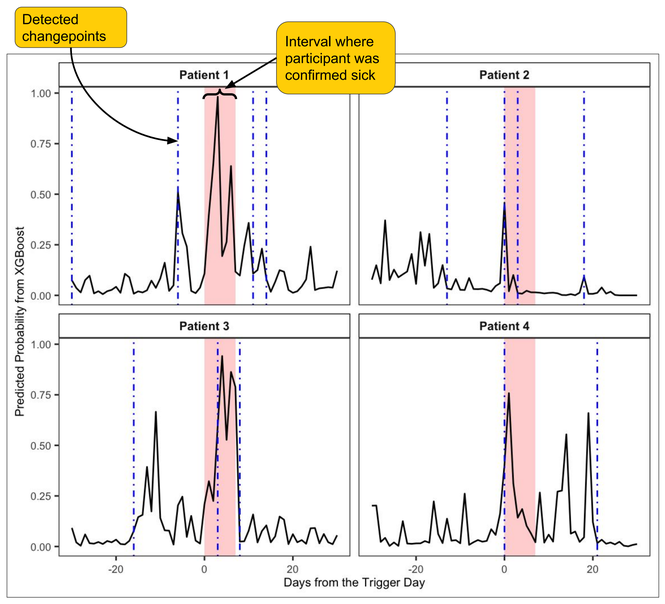

Fig 5. Predicted probability of a patient having respiratory viral infections 30 days before and after the testing date. Changepoints in predicted probability were detected (blue dashed lines) and compared to positive period (red regions).

Summary

- XGBoost (AUC-ROC = 0.788) has a better performance than DNN (AUC-ROC = 0.747), so we applied changepoint detection to the predicted probabilities obtained from XGBoost.
- In some cases, the changepoint was detected a few days earlier or later (fig 5, patient 1). The onset point was systematically created so there may be potential inaccuracies. Therefore, the lag could reveal the true onset point according to their bodily changes.
- Despite sometimes missing the onset point, able to occasionally identify the recovery point (fig 5, patient 3).

Future Work

Potential expansions of this project:

- Improved or different predictors for diagnosing illness status
- Investigation into handling missing data
- Observing the correlation between predictors
- Exploring UMAP and TSNE in modeling or variable importance/selection (Fig 6, Fig 7)

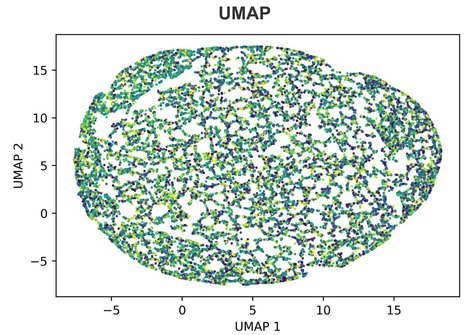

Fig 6. UMAP plot of with Euclidean minimum distance of 0.1.

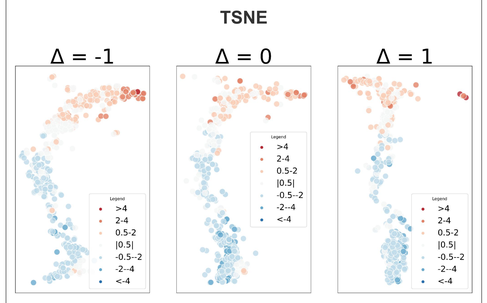

Fig 7. t-SNE plot of activity calories with perplexity 40.

References

[1] Alavi, A., Bogu, G.K., Wang, M. *et al.* Real-time alerting system for COVID-19 and other stress events using wearable data. *Nat Med* 28, 175-184 (2022).

[2] Merrill, M.A. and Althoff T. Self-supervised pretraining and transfer learning enable flu and COVID-19 predictions in small mobile sensing datasets. *arXiv:2205.13607* (2022).

[3] Kolbeinsson, A., Gade P., Kainkaryam R. *et al.* Self-supervision of wearable sensors time-series data for influenza detection. *arXiv: 2112.13755* (2021).

[4] Kotnik, J.H., Cooper, S., Smedinghoff, S. *et al.* 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. J.Clin Microbiol 60(3): e0207021. *doi: 10.1128/JCM.02070-21* (2022).

[5] Wilson, S. The Mice Algorithm (2021).

[6] Monaghan, C.K., Larkin, J.W., Chaudhuri, S. *et al.* Machine learning for prediction of patients on hemodialysis with an undetected SARS-CoV-2 infection. Kidney 360 2(3):p 456-468. *doi: 10.34067/KID.0003802020* (2021).