HARNESSING PHYSIOLOGICAL DATA FOR EARLY WARNING OF THREAT EXPOSURE

Model-agnostic Framework For Evaluating Performance, Robustness, And Fairness Of Machine Learning Infection Detection

Consuelo Cuevas MIT Lincoln Laboratory **Thomas Koker** MIT Lincoln Laboratory **Shakti Davis** MIT Lincoln Laboratory
Natalie Damaso MIT Lincoln Laboratory **Kajal Claypool** MIT Lincoln Laboratory

The Defense Threat Reduction Agency (DTRA) funds several efforts to develop artificial intelligence/ machine learning (AI/ML) algorithms to provide early warning of biological and chemical exposure based on physiological measurements obtained from wearable devices. AI/ ML models have the desirable property of learning subtle patterns from data; however, this can lead to unintended consequences and the resulting models are often considered “black boxes” since the underlying factors that led to a decision are difficult to decipher and explain. To address this concern, we propose a model-agnostic framework for evaluating wearables-based infection detection models that probes three primary areas: general performance, robustness, and fairness.

General algorithm performance is measured in this framework using receiver operating characteristic (ROC) curves and area under the curve (AUC) scores across stratified cross-validation folds and at one-or-more times of interest for a chemical or biological response. In addition, given the temporal nature of physiological responses, visualizations of the model’s sensitivity and specificity over time are also evaluated. To measure model robustness against feature inaccuracies, temporal gaps in the data, or a subset of available features from a wearable device, model’s performance metrics are re-evaluated with degraded input data and different degradation conditions. Lastly, to understand fairness of the model, the above evaluation and robustness metrics are repeated on sub-populations of the study cohort to probe for potential performance discrepancies among underrepresented demographic groups such as females or individuals with comorbidities.

We demonstrate the value of this framework for comparing and contrasting different models with an example of COVID-19 detection based on the Rapid Analysis of Threat Exposure (RATE) dataset. The RATE dataset was created by a large-scale study data of COVID-19 monitoring from the United States Department of Defense personnel wearing commercial wearable devices. Using physiological measurements from this dataset, we evaluate COVID-19 detection models under two disparate architectures: a classical ML model using random forests and a deep learning neural network model using long short-term memory (LSTM). We use our framework to assess algorithm performance, robustness and fairness at different stages of infection (prodromal and acute illness), robustness to measurement errors and missing features, and fairness of the models on the basis of gender and age group. For instance, preliminary results for the random forest model show a 0.75 AUC over a 3-day window before and after the diagnosis date across the RATE population. However, a deeper breakdown shows that there is an 8% difference in AUC performance across genders indicating that AUC individually does not represent the entirety of the model’s performance.

Understanding and measuring AI/ML model performance is complicated and requires a careful and thoughtful approach. The proposed framework promotes opportunities for directly comparing results across disparate AI/ML models, improves the understanding of the model behavior under degraded conditions and probes for potentially harmful biases in the trained model that might otherwise go unnoticed. As work continues in the integrated early warning area, this effort provides a framework to evaluate these models in a consistent manner.

This material is based upon DTRA funding CB10522 supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702 -15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.