

## RADAR-Pipeline: Scalable Feature Generation for Mobile Health Data

Heet Sankesara<sup>1</sup>, Yatharth Ranjan<sup>1</sup>, Pauline Conde<sup>1</sup>, Zulqarnain Zulqarnain<sup>1</sup>, Akash Roy Choudhury<sup>1</sup>, and Amos Folarin<sup>1</sup>

<sup>1</sup>The Department of Biostatistics and Health informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

### Introduction & Background

RADAR-Pipeline is an open-source Python framework designed to simplify and enhance mobile health data analysis. It has been designed to efficiently read and process the large amount of data generated through the RADAR-Base platform. RADAR-base is a scalable, real-time streaming and analytics open-source platform to facilitate research access and customisation requirements. Studies using the Radar-base platform have collected fine-grained longitudinal data from wearables and phones. The data can potentially create multitudes of digital biomarkers, which can be used to inform us greatly about the disease condition. Due to the sheer size of the data, it can be difficult for researchers to read and process those data – a common task is identifying useful features and common data processing/analysis steps previously used by the community. Up to now, these have been hand-crafted by individual data scientists, often lacking the capability to be easily reused by the community without author-specific knowledge.

Furthermore, generating variables based on already established research on a larger scale can be challenging and could hinder replication. Hence, we have designed RADAR-Pipeline to help researchers overcome these challenges. It empowers them to create and share their data analysis and visualisation pipelines, fostering collaboration and knowledge sharing within the research community.

### Objectives & Approach

The primary objective of RADAR-Pipeline is to offer researchers a user-friendly and powerful platform to develop and share their research. Researchers can build reusable analysis and visualisation pipelines to ensure consistent and reliable results. It simplifies big data analysis by leveraging Apache Spark to handle large and complex mobile health datasets efficiently. Researchers can also save time and effort by reusing and extending existing pipelines built by others. Finally, the RADAR-Pipeline promotes collaboration and recognition by allowing researchers to share their work through the RADAR-base Analytics Catalogue, making their pipelines citable and accessible to the wider research community.

Whilst Radar-pipeline has been designed to read data from Radar-base, it can also be used to read data from any dataset which uses Hadoop Distributed File System (HDFS) file system namespace.

### Relevance to Digital Footprints

Mobile health data is rich and valuable for understanding human behaviour and health. RADAR-Pipeline addresses the challenges associated with analysing large and complex mobile health datasets, enabling researchers to extract valuable insights that can be used to (1) Improve public health: By enabling efficient analysis of large-scale mobile health data, RADAR-Pipeline can contribute to research efforts aimed at improving population health outcomes and developing effective interventions; (2) Personalised healthcare: By facilitating the extraction of individual-level features from mobile health data, RADAR-Pipeline can seamlessly be integrated with Kafka data streams and machine learning pipelines to process the data in real-time, which can then be utilised to create more effective and targeted real-time interventions. (3) Promote reproducible research: The framework's emphasis on transparency and reproducibility in research aligns with the conference's focus on the responsible use of digital mobile health data.

### Conclusions & Implications

RADAR-Pipeline is a valuable tool for researchers, offering them the means to harness the potential of mobile health data. By adopting this framework, researchers can achieve efficient and scalable data analysis, thereby streamlining the extracting insights from digital footprints. This efficiency enables researchers to delve deeper into the data and uncover valuable patterns and trends.

Furthermore, RADAR-Pipeline promotes collaboration and knowledge sharing within the research community. By providing a standardised framework for data analysis, RADAR-Pipeline facilitates collaboration among researchers, leading to the sharing of best practices and the dissemination of knowledge.

