Advancing Machine Learning Systems for Temporal Healthcare Data

Context and Motivation

Healthcare generates vast amounts of temporal and spatiotemporal data from wearable sensors, electronic health records (EHR), and infectious disease epidemiology. Traditional statistical models often struggle to capture the complex temporal and spatial dependencies inherent in this data. However, recent advancements in machine learning (ML), particularly in deep learning and graph-based models, present transformative opportunities for enhancing predictive accuracy, patient monitoring, clinical outcome prediction, and early disease detection.

This project seeks to develop advanced ML frameworks to effectively model and analyze healthcare time series data, leveraging spatial dependencies to enhance diagnostic and prognostic capabilities. Given the inherent noise and uncertainty in biological and medical time series data, the primary objective is to design interpretable, robust, and scalable ML systems. By integrating domain-specific medical knowledge with cutting-edge ML techniques, we aim to improve the reliability and transparency of predictive models, fostering the development of uncertainty-aware decision-making frameworks for healthcare applications.

Scientific Objectives

This research aims to develop advanced machine learning systems that address the challenges posed by irregularities in medical data, with a particular focus on healthcare applications such as epidemic spread, imbalanced medical data classification, and spatiotemporal disease modeling. Our objective is to design and analyze machine learning models that are scalable, interpretable, robust, and uncertainty aware. We will explore the following key research directions:

1. *Imbalanced Univariate Time Series Classification:* Electrocardiograms (ECGs) are essential tools in diagnosing heart diseases, as they provide time series data reflecting heart rate patterns. Healthy individuals typically exhibit regular, rhythmic heart rates, while patients with cardiovascular conditions often show irregular rhythms or elevated heart rates. The task is to classify these patterns, enabling accurate health status assessments. Given the inherent imbalance in medical datasets, where some classes are underrepresented, our goal is to develop robust imbalanced classifiers for time series data, ensuring reliability across diverse patient populations.
2. *Multivariate Spatiotemporal Data Classification:* Healthcare data (e.g., ECGs) often includes multiple time series from various health sensors, such as blood pressure monitors, glucose sensors, and pulse oximeters. These time series represent different aspects of a patient's health and analyzing them collectively offers insights into interrelated health patterns. We aim to develop spatiotemporal graph deep learning-based classification model that can capture both temporal and spatial dependencies across these variables. The primary objective is to transform high-dimensional, complex time series data into more comprehensible representations, facilitating accurate classification and analysis.
3. *Epidemic-Guided Machine Learning Models for Infectious Disease Forecasting:* Infectious diseases (e.g., Measles, Tuberculosis) remain major contributors to global morbidity and mortality, often following epidemic patterns. Instead of solely relying on data-driven models, incorporating disease-specific knowledge and epidemic priors into deep learning systems can improve predictive accuracy. This research aims to develop epidemic-informed deep learning model to forecast epidemic outbreaks, enhancing our understanding of disease dynamics and enabling more effective public health responses.

4. *Forecasting Spatiotemporal Epidemiological Data:* Traditional epidemic models typically focus on the temporal dimension, neglecting spatial variation. Recently, deep learning-based spatial-temporal models have shown promise, but challenges remain, especially when surveillance data is noisy or sparse. This research seeks to integrate structural, spatiotemporal, and epidemiological data within a unified framework, overcoming existing methodological limitations to improve the forecasting of epidemic dynamics.

Experiments and Validation

Apart from theoretical developments, our objective is to integrate these models into wearable monitoring technologies and clinical decision-making tools, thus enabling more personalized, effective, and preventive patient care. ICAN will provide large-scale clinical datasets to validate the data-driven methods to industrial case studies such as heart rate adaption (chronotropic response) in heart failure patients.

Year-wise plan

Year 1: A comprehensive literature review (both from the methodological side and application perspective) will guide the methodology and contextualize the findings. The focus will also be on curating and preprocessing the existing datasets for meeting the scientific objectives. Advanced machine learning models will be developed and refined to improve the prediction of chronotropic response anomalies to meet the first objectives.

Year 2: Developing advanced ML techniques integrated with time series feature selection, constructing imbalanced temporal data classifiers, and building multivariate techniques, will be part of the innovation to study during this time. A significant amount of time will be dedicated to spatiotemporal graph neural networks designed to forecast epidemics with disease prior.

Year 3: In the final year, the project will validate the biomarker's diagnostic performance in clinical settings, ensuring its robustness across diverse patient populations. Efforts will also focus on developing a prototype wearable device and epidemic decision support systems for real-time monitoring, translating research findings into practical tools for clinicians and patients. Results will be disseminated through publications, conference presentations, and the final dissertation.

Candidate Profile. Candidates should have: (1) Strong Statistics, Forecasting, and ML background, (2) Interest in health applications, (3) Strong programming skills in Python and R, and (4) Ability to communicate effectively in English, both orally and in writing.

Team and Location. The candidate must physically study at Sorbonne University Abu Dhabi (UAE) for the duration of their PhD, although their PhD enrollment (and eventual degree) will be from Sorbonne University (Paris, France). The primary adviser will be Dr. Tanujit Chakraborty, co-advised by Dr. Maharajah Ponnaiah (ICAN). Pr. Gérard Biau will sign as HDR holder.

Key References:

[1] Jin, Ming, et al. "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[2] Rodríguez, Alexander, et al. "Machine learning for data-centric epidemic forecasting." Nature Machine Intelligence 6.10 (2024): 1122-1131.
[3] Panja, Madhurima, Chakraborty, Tanujit, Kumar, Uttam and Liu, Nan. "Epicasting: an ensemble wavelet neural network for forecasting epidemics." Neural Networks 165 (2023): 185-212.
[4] Barman, Madhab, Panja, Madhurima, Mishra, Nachiketa, and Chakraborty, Tanujit. "Epidemic-guided deep learning for spatiotemporal forecasting of Tuberculosis outbreak." *arXiv preprint arXiv:2502.10786* (2025).