PROJECT 5 Symptom-Onset prediction with LSTM and sliding window

Business scenario
Wearables capture continuous vitals; clinicians need to spot the *day* a symptom becomes acute.

Dataset

Synthetic dataset:

(or create your own)

- **Rows:** 2,000 (50 patients × 40 days)
- **Columns**

name	description
patient_id	integer ID (1 – 50)
day	timeline index (0 – 39)
temperature_C	body-temperature (°C) with realistic drift/noise
heart_rate_bpm	heart-rate (beats / min)
sbp_mmHg	systolic blood-pressure (mm Hg)
label	**CHRONIC**, **ACUTE_ONSET**, or **UNKNOWN** (for 10 % randomly missing days)

- **Temporal pattern:**
  - A patient-specific **acute-onset day** is chosen between day 7 – 30.
  - Vitals **gradually rise** toward this day (sequence context ⇒ LSTM advantage).
  - On the onset day they **spike**, then stabilise at an elevated plateau.

Use sliding windows (e.g., 7–10 days) to show how an LSTM detects the ramp-up trend and predicts the ACUTE_ONSET tag earlier and more accurately than a per-day logistic baseline.

Core technique
A unidirectional or bidirectional **LSTM** that outputs a 3-class tag (acute / chronic / unknown) for each day in a n-day window.

Key steps

- Create sliding windows; apply label smoothing for borderline days.
- Train the LSTM; compare to a logistic-regression baseline.
- Plot per-timestep predictions and ground truth.

Deliverables

- Notebook with exploratory data analysis (EDA).
- Training curves and confusion matrices.
- Brief write-up of error patterns (e.g., weekend gaps).
- **Per-timestep F1** and **AUROC** on the test split.
- Comparative table: LSTM vs. baseline.
- Discussion of false-positive clusters.

Labeled targets.
You supply each day (or time step) in the 30-day window with one of three explicit labels:

- ACUTE_ONSET * | * CHRONIC * | * UNKNOWN *.
  The model learns to map an input vector of vitals for that day (possibly including the preceding context) to the ground-truth tag.

Cross-entropy loss.
Training minimises multi-class cross-entropy between the predicted soft-max distribution and the one-hot label vector, which is the hallmark of supervised classification.

What "spot the day a symptom becomes acute" really means

1. **Granularity** – You have a *time series* of daily (or hourly) vital-sign vectors for one patient over a fixed horizon (e.g., the first 30 days after surgery).
2. **Label definition** – For each day *t* you want to predict whether, **on that day**, the patient crossed the clinical threshold that converts a lingering complaint ("chronic/ongoing") into an **acute episode** that deserves new intervention.

3. **Sequence dependence –** The judgment for day *t* is rarely based on its measurements in isolation; clinicians look for *trends* (e.g., a steady rise in temperature plus a sudden CRP spike over the previous three days).

Why LSTM is proposed?

Aspect	**Logistic-regression baseline**	**LSTM / sequence labeller**
Input representation	Typically a *single-day* feature vector (or a hand-crafted Δ-vector such as "current minus previous day").	A *window* of consecutive daily vectors, preserving their order.
Temporal modelling	No memory – treats each day independently unless you manually add lag features.	Hidden state carries forward information about the pattern up to day *t*.
Capturing trends / change points	Must be encoded explicitly (e.g., "3-day moving average > threshold").	Learns patterns like "gradual climb then sudden jump" automatically.
Pedagogical value	Shows limitations of static models on dynamic data (good baseline).	Gives students hands-on practice with sequence tagging, padding, masking, imbalance handling.

So the expectation is **not** merely a three-class logistic classifier on current-day vitals; it is a *sequence-aware* classification where the model can "see" earlier readings and learn temporal cues without you specifying them.

How to structure it in practice

1. **Sliding-window setup**
   - For every patient, extract overlapping windows (e.g., days 1-7, 2-8, ...).
   - The model receives the full 7-day sequence and outputs a tag for the **final** day (or for every day, if you prefer a true sequence-to-sequence labeller).
2. **Label creation**
   - ACUTE_ONSET = first day that meets clinical escalation criteria.

- CHRONIC = earlier days where symptom is present but stable.
- UNKNOWN = missing data or clearly asymptomatic periods.

3. **Baselines to compare**
   - **Logistic regression** on the *current-day* vector (plus optional lagged deltas) → shows what you lose when you ignore history.
   - **LSTM** (or GRU, 1-D CNN with dilated filters) → expected to pick up gradual drifts or sudden jumps.
   - (Optional) **Change-point detector** using reconstruction error → unsupervised variant if labels are sparse.
4. **Metrics**
   - Per-day F1 and AUROC.
   - Detection lag: average number of days between true acute onset and first correct positive prediction.
