MAIS 202 – Project Proposal
Title: Early Sepsis Risk Prediction

Team: Vipul Modi, Tanav Bansal

1) Choice of dataset
We will use **[Dataset: PhysioNet 2019 Sepsis Challenge]**, which provides **hourly time-series** of vitals and labs (e.g., heart rate, mean arterial pressure, respiratory rate, temperature, WBC, lactate, creatinine) and a timestamp for **sepsis onset**. These research-grade datasets are large, standardized, and curated for early-warning tasks, enabling reproducible training/evaluation without clinical access. **Educational demo only; not a medical device.**

2) Methodology (high-level plan)
(a) Data preprocessing. Resample to **hourly**; apply physiologic caps; forward-fill with carry limits; add **missingness indicators**. Engineer features over **1/6/12h** windows (deltas, means, mins/maxes, slopes), time-since-admission, and simple treatment flags if present. Split **by patient** into train/val/test (no leakage).
(b) Models. Baseline: LightGBM with class weighting + early stopping. **Stretch:** GRU or Temporal Fusion Transformer with masking; focal loss for rare positives. **Calibration:** temperature scaling so 0.18 ≈ **18%** risk. **Explainability:** SHAP (global + per-patient) and partial-dependence for key vitals.
(c) Evaluation & feasibility. Primary: AUROC, **AUPRC** (imbalance), Expected Calibration Error (ECE), reliability plots. **Operational: Sensitivity at fixed FPR** (e.g., 20–30%) and **Lead Time** (avg hours before recorded onset when an alert would fire). **Baselines:** majority class, SOFA-like rule, and "last-value only" logistic regression. **Error analysis:** subgroup slices (age/sex/ICU type) where available.

3) Application (webapp)
An **educational demo** that visualizes risk over time with transparent reasons.

- **Input:** choose example patient or upload CSV of hourly vitals/labs.
- **Output:** (1) 6-hour risk timeline, (2) adjustable alert threshold with **alerts/day** estimate (alert burden), (3) top contributing features per timestamp (SHAP), (4) simple "what-if" sliders (e.g., **+5 MAP → risk ↓x%**).
- **Tech:** Streamlit or React+FastAPI; saved model (pickle/onnx); small sample cohorts bundled.

4) Baselines, metrics, targets
Baselines: majority class; SOFA-like heuristic; last-value logistic regression.
Targets (dataset-dependent): AUPRC ≥ 0.30–0.35, **AUROC ≥ 0.80**, **mean Lead Time ≥ 4h** at ~25% FPR, **ECE ≤ 0.05**. Include a **threshold trade-off table**: Sensitivity / PPV / FPR and **alerts per 24h**.

5) Risks & mitigations
Label variability → follow dataset's definition; document limits.
Alert fatigue → calibrated probabilities; evaluate at fixed FPR + show alerts/day.
Overfitting → strict patient-level split; (stretch) external validation on a second cohort.
Missing/noisy data → missingness flags; robust windowed trends.

6) Timeline & repo
Week 1: EDA, preprocessing, LightGBM baseline, initial metrics.
Week 2: Calibration + SHAP; lead-time & alert-burden analysis; Streamlit demo.
Week 3 (stretch): GRU/TFT; external validation; counterfactual UI.