POTENTIAL RESEARCH GAPS I PICKED

5. Temporal Irregularities in EHR Data

Problem:

EHR data is irregularly sampled—vitals and labs are recorded at inconsistent intervals depending on clinical need, not model requirements.

Why it matters:

Most ML models (like standard LSTMs or XGBoost) assume uniform time steps, which can distort the actual clinical progression of sepsis.

Promising Approaches:

- Time-Aware LSTM (T-LSTM): Incorporates time gaps between events as part of the model input.
- GRU-D: Modified GRU architecture that learns to decay older observations depending on time gaps.
- Neural ODEs / Latent ODEs: Model continuous time-series with irregular intervals.

6. Early vs. Late Sepsis Detection

Problem:

Many models predict sepsis too close to the onset (e.g., 1–2 hours before), which doesn't give clinicians enough time to intervene.

Why it matters:

The clinical utility of a model is significantly higher if it can predict 6–12 hours in advance while maintaining reasonable accuracy.

Research Strategy:

- Design your model with time-to-onset labels (e.g., predict sepsis at least X hours ahead).
- Evaluate prediction window performance trade-offs (e.g., accuracy vs. lead time).
- Implement sliding window or multi-horizon forecasting.

9. Multimodal Data Integration

Problem:

Most models only use structured data (vitals, labs). Rich insights are ignored in clinical notes, medication records, and imaging.

Why it matters:

Sepsis is complex and its onset may be indicated by multiple data types. Fusing structured and unstructured data can improve performance.

Solutions/Approaches:

- BERT for Clinical Notes (e.g., BioBERT, ClinicalBERT) combined with vitals.
- Late fusion models: Separate sub-models for each modality, merged at decision level.
- Attention-based fusion: Dynamically weigh modalities based on importance.
- Use multimodal transformers for joint learning.

SOLVING AND IMPLEMENTATION

Phase 1: Data Preparation (Foundation for All Gaps)

1. Dataset Selection

Use a dataset like:

- MIMIC-III or MIMIC-IV (public ICU EHR dataset)
- Includes vitals, labs, demographics, and clinical notes (multimodal)

2. Preprocessing Steps

- Align all data to patient stay ID and timestamps
- Convert vitals and labs into time-series (hourly binning or fixed intervals)
- Identify and label sepsis onset time using Sepsis-3 definitions
- For clinical notes: Clean text (remove PHI, headers, etc.)

Phase 2: Temporal Irregularity Handling (Gap 5)

A. Problem

Vitals/labs are sampled at irregular intervals.

B. Solution Implementation

- GRU-D or Time-Aware LSTM:
- Use time gaps between observations as input
- Tools: PyTorch or TensorFlow custom layers
- Alternative: Neural ODEs
- Use torchdiffeq library in PyTorch
- Encode time-dependent latent dynamics

C. Steps

1. Construct a data matrix with timestamps and time-gaps.
2. Train GRU-D or T-LSTM on this matrix.
3. Compare with standard LSTM to demonstrate improvement.

Phase 3: Early Detection Optimization (Gap 6)

A. Problem

Models detect sepsis too close to onset.

B. Solution Implementation

- Create labels for predicting 6, 12, 24 hours before onset (label shifting)
- Use multi-horizon forecasting to assess early predictions
- Monitor trade-offs between precision vs. lead time

C. Steps

1. Shift sepsis onset label backwards in time (6/12/24 hrs).
2. Train models with sliding windows or sequence-to-one prediction.
3. Evaluate:

- AUROC
- Sensitivity @ different prediction windows

Phase 4: Multimodal Integration (Gap 9)

A. Problem

Only structured data is used; text (clinical notes) is ignored.

B. Solution Implementation

- Use ClinicalBERT or BioBERT to encode text notes
- Combine embeddings with vitals using late fusion or attention fusion

C. Steps

1. Extract discharge summaries and notes near sepsis onset.
2. Tokenize and encode using transformers library (Hugging Face).
3. Concatenate BERT embeddings with time-series features.
4. Train either:

- Late Fusion: Separate models for text and vitals, then merge
- Joint Model: Attention or multimodal transformer (advanced)

Phase 5: Final Model Pipeline

1. Input: Vitals + Labs (irregular time-series) + Clinical Notes
2. Time-aware LSTM/GRU-D processes time-series
3. BERT model processes notes
4. Attention mechanism or fusion layer combines features
5. Output: Sepsis risk score with timestamp

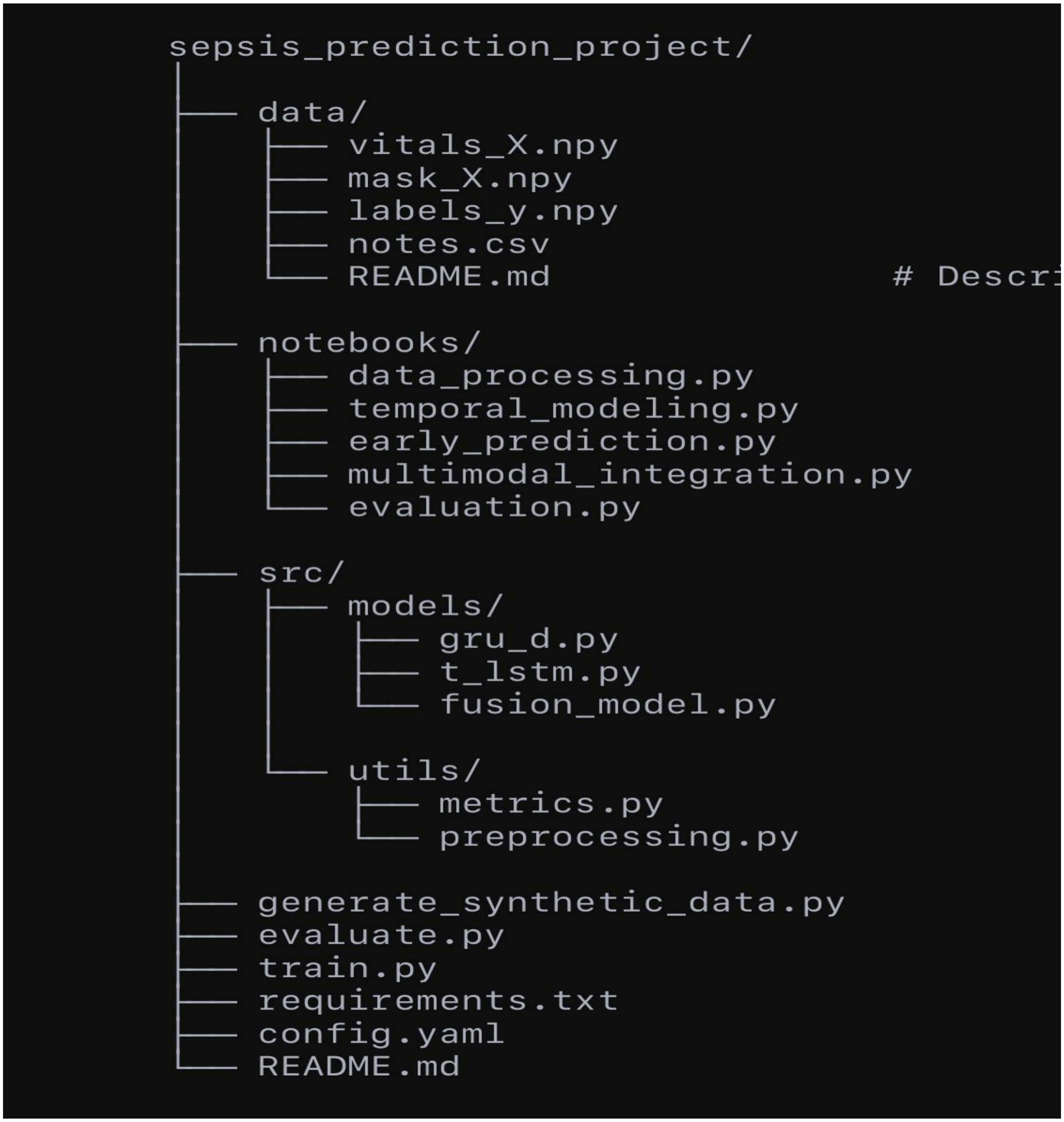
Phase 6: Evaluation and Validation

- Metrics: AUROC, AUPRC, Sensitivity, Specificity, Lead Time
- Ablation Study:
- Without time irregularity handling
- Without text
- Without early prediction shift
- Comparison with baselines: SIRS, qSOFA, NEWS

Tech Stack Suggestions

Component	Tool / Library
Time-Series Models	PyTorch, TensorFlow
NLP/BERT Embeddings	Hugging Face transformers
Data Preprocessing	Pandas, NumPy
Visualization	Seaborn, Matplotlib
Data Source	MIMIC-III / IV (officially) [un-officially I've added create_synthetic_data.py in the root structure, which creates/augments/synthesis around 400 patient data as required by the modals]

The below will be the directory structure



```
sepsis_prediction_project/
│
├── data/
│   ├── vitals_X.npy
│   ├── mask_X.npy
│   ├── labels_y.npy
│   ├── notes.csv
│   └── README.md               # Descri
│
├── notebooks/
│   ├── data_processing.py
│   ├── temporal_modeling.py
│   ├── early_prediction.py
│   ├── multimodal_integration.py
│   └── evaluation.py
│
├── src/
│   ├── models/
│   │   ├── gru_d.py
│   │   ├── t_lstm.py
│   │   └── fusion_model.py
│   │
│   └── utils/
│       ├── metrics.py
│       └── preprocessing.py
│
├── generate_synthetic_data.py
├── evaluate.py
├── train.py
├── requirements.txt
├── config.yaml
└── README.md
```