Habib University

Computer Vision

Project Proposal

“Multilevel Contrastive Learning for Enhanced Clinical Prediction in Electronic Health Records”

Name:	Ahtisham Uddin	(au08429)
	Bilal Ahmed	(ba08018)
	Manqad Raza	(mr08456)
Class:	2026	
Instructor:	Muhammad Farhan	
Submission Date:	6th December 2024	

Contents

1 Introduction

1.1 Motivation

Advancements in artificial intelligence are transforming healthcare, with Electronic Health Records (EHRs) serving as critical sources of data for clinical predictions. Despite this progress, many models treat medical measurements in isolation, missing the context provided by grouping features based on clinical relationships, such as organ systems or diagnostic contexts. Furthermore, EHRs inherently capture patient health trajectories over time, which is often overlooked in traditional approaches. Contrastive Learning (CL) has emerged as a promising solution to address these gaps, offering a way to capture complex relationships within multivariate and time-dependent medical data. This project is motivated by the need to combine feature grouping with CL methods, aiming to develop a model that can better capture the multidimensional nature of EHRs for improved patient monitoring and prediction

1.2 Significance of the Project

This project seeks to enhance clinical decision-making by developing a model that leverages contrastive learning across clinically meaningful feature groups. By capturing hierarchical relationships within medical data, such a model can improve the adaptability of predictive algorithms across different healthcare settings, from intensive care to general inpatient monitoring. Additionally, this approach aligns with the goal of personalized medicine by focusing on individual patient patterns and temporal dependencies. The outcomes of this project have the potential to bridge academic research and practical healthcare applications, enabling more reliable and interpretable clinical predictions.

1.3 Description of the Project

This project will integrate contrastive learning techniques with feature grouping to improve medical time-series analysis. Specifically, we will create positive and negative pairs at different levels: feature groups, temporal neighborhoods, and patient-specific data. The proposed contrastive learning objective aims to capture both high-level and granular patterns within EHR data, enhancing model performance on clinical prediction tasks. The project will involve a literature review, model development, and rigorous evaluation to assess improvements in metrics such as accuracy and sensitivity. Through this approach, we aim to develop a more interpretable and adaptable model for clinical prediction tasks in healthcare.

2 Methodology

2.1 Design/Methodology Phase

In the design phase, our group will begin by creating a straightforward framework to incorporate contrastive learning for analyzing EHR time-series data. We will start with a focused literature review to understand the basics of self-supervised learning, contrastive learning techniques, and how hierarchical feature grouping has been applied in medical data. The main goal of this phase is to define simple positive and negative pair structures at different levels—feature group, temporal neighborhood, and patient level—so that the model can capture essential relationships within the data.

Next, we will outline the key data preprocessing steps needed for working with EHR data. This includes aligning time-series data, normalizing features, and addressing missing values. We will also group features based on medical categories, like organ systems, to serve as a foundation for creating pairs in the contrastive learning process. The output of this phase will be a basic outline of the model structure and training approach.

2.2 Implementation Phase

This phase involves building the initial model, with a focus on simplicity to ensure we can complete it within three weeks. Key steps include:

- **Data Preprocessing and Feature Grouping:** Our team will implement preprocessing steps such as normalizing the data and handling missing values, using feature groupings based on clinical relevance. This will form the basis for generating contrastive pairs in the learning process.

- **Contrastive Learning Framework:** We will start with a basic version of the contrastive learning framework, creating positive and negative pairs based on one or two levels of granularity (likely temporal and feature group). This approach will help us capture relationships within EHR data without making the model overly complex.

- **Model Architecture:** To keep things manageable, we will use a simple neural network architecture suitable for time-series data, such as an LSTM. This choice allows us to focus on applying the contrastive learning objective without adding unnecessary complexity.

- **Training Pipeline:** We will set up a basic training pipeline with a contrastive loss function. To monitor progress and avoid overfitting, we'll incorporate validation techniques and, if possible, use GPU resources for faster training.

2.3 Testing and Evaluation Phase

In the testing and evaluation phase, we will focus on assessing how well the model performs on a few specific tasks and analyzing the impact of the contrastive learning approach. Key steps include:

- **Basic Benchmark Testing:** We will evaluate the model on a straightforward prediction task, such as mortality prediction or length-of-stay prediction, using common metrics like accuracy. This will provide a baseline measure of the model's performance.

- **Simple Ablation Study:** To understand the importance of each granularity level (e.g., temporal vs. feature group), we will disable one level at a time and observe any changes in model performance. This analysis will help identify which parts of the model contribute most to its effectiveness.

- **Comparison with a Baseline:** We will compare our model's performance with a simple supervised baseline model. This comparison will allow us to see if adding contrastive learning actually improves results for our tasks.

- **Interpretability Check:** Finally, we will examine the model's embeddings to understand what it's learning, especially in terms of feature group relationships. We will also conduct an error analysis to identify common mistakes, using these insights to suggest areas for future improvement.

3 Characteristic Features

This project combines several distinctive features that make it unique within the field of EHR-based predictive modeling. The key characteristics include:

- **Multilevel Contrastive Learning Approach:** A primary feature of this project is the use of a contrastive learning framework across multiple levels, allowing the model to capture patterns at different levels of detail. By generating positive and negative pairs at the feature group, temporal, and patient-specific levels, the model will be better equipped to understand complex patterns in medical data.

- **Clinical Feature Grouping:** The project will incorporate clinical knowledge to group features hierarchically, such as by organ systems or types of measurements. This structure helps the model interpret relationships within clinical data more effectively, making it more relevant and interpretable for medical applications.

- **Temporal and Patient Contexts:** Beyond simple feature grouping, this project includes both time-based and patient-specific contexts in the model. This allows the model to understand trends in patient health over time, which is particularly important in clinical settings where patient conditions change gradually or rapidly depending on circumstances.

- **Adaptability to Diverse Clinical Data:** By using contrastive learning, the model aims to handle the variety in clinical data across different healthcare settings, patient demographics, and measurement methods. This adaptability is important for models used in real-world healthcare, where data can vary significantly.

- **Emphasis on Interpretability:** The model is designed with interpretability in mind, aiming to show how different feature groups and patient-specific factors influence predictions. This transparency is essential in clinical environments, where healthcare providers require models that offer clear insights into patient health.

4 Project Planning

The project is organized into a structured timeline over the three-week period, with specific goals for each phase to ensure timely progress.

Week 1: Research and Design

- **Literature Review:** Research relevant topics, including contrastive learning and feature grouping in medical data.
- **Define Model Structure:** Decide on a simple architecture and identify the levels of granularity (feature group, temporal, patient) for contrastive learning.
- **Data Exploration:** Familiarize the team with the EHR dataset, identifying key preprocessing steps needed for implementation.

Week 2: Implementation

- **Preprocessing and Feature Grouping:** Implement data preprocessing steps and define feature groupings.
- **Contrastive Pair Generation:** Start creating positive and negative pairs based on the chosen granularities.
- **Model Development:** Implement a basic model architecture, likely an LSTM or similar time-series model, and set up the contrastive learning framework.

Week 3: Testing and Evaluation

- **Performance Evaluation:** Test the model on a basic prediction task, assessing its performance with metrics like accuracy and sensitivity.
- **Ablation Testing:** Conduct simple ablation tests to analyze the impact of each contrastive learning level.
- **Comparison with Baseline:** Compare the results with a baseline model to evaluate the benefits of the contrastive learning approach.
- **Interpretability Analysis:** Examine model embeddings to understand how it's learning relationships in the data.

Final Report and Presentation

- **Documentation:** Compile findings, methodologies, and results into a final report.
- **Prepare Presentation:** Create a presentation summarizing the project's objectives, methods, results, and conclusions.

5 Required Hardware and Software

To handle the computational demands of training a deep learning model on EHR time-series data for contrastive learning, the recommended hardware includes:

5.1 Hardware Requirements

- **GPU (Optional):** 4GB VRAM or higher to accelerate model training.
- **Multi-core CPU:** Recent multi-core processor for efficient data processing.
- **RAM:** Minimum 8GB; 16GB recommended for smoother performance.
- **Storage:** 100GB SSD preferred for faster data access.

5.2 Software Requirements

The following software tools and libraries will be used to implement and evaluate contrastive learning on EHR data:

- Python – The primary language for all coding and data processing tasks.
- PyTorch or TensorFlow – To build and train the contrastive learning model, with PyTorch preferred for its ease of use and extensive documentation.
- Pandas and NumPy – For handling and processing EHR data.
- Matplotlib and Seaborn – To visualize data trends, model performance, and embeddings.
- Git – For version control and collaboration among team members.
- Jupyter Notebook or VS Code – For interactive coding, testing, and sharing results. Jupyter is useful for data exploration, while VS Code supports efficient project management.

6 References

The following are key references that provide foundational concepts, methodologies, and recent advances relevant to this project:

- Imrie, F., Bradley, C., van der Schaar, M., & Rashbass, J. (2022). *Understanding Feature Groupings in Clinical Time-Series Prediction Models.* Journal of Biomedical Informatics, 128, 104003.
- Kelly, M., & Semsarion, A. (2009). *Predefined Feature Groupings in Clinical Data: An Approach to Enhanced Interpretability and Prediction.* Proceedings of the Annual Symposium on Biomedical Data.
- Yeche, H., Antognini, D., Mosko, J. D., & Pal, C. (2021). *Contrastive Learning in ICU Patient State Prediction: A Study on Domain Heterogeneity.* ICML Workshop on Health Intelligence, 12-20.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations.* In Proceedings of the 37th International Conference on Machine Learning (ICML).