CSE 6250 Big Data for Healthcare
Spring 2025
Project Proposal - Deliverable 2 - 17th Feb 2025
Team C3

Nikhil Kapila nkapila6@gatech.edu
Tejas Rathi trathi9@gatech.edu

1. Paper 1:

- **Index-Title-Author**: 07-AI-Driven Clinical Decision Support: Enhancing Disease Diagnosis Exploiting Patients Similarity by Carmela Comito, Deborah Falcone, Agostino Forestiero
- **Task**: Develop an AI-driven clinical decision support system (CDS) by using patient similarity scores on symptoms and preliminary diagnoses.
- **Innovation**: Patient similarity-based diagnosis prediction that considers a single diseas & system predicts multiple medical conditions, incorporates multiple data sources expanding beyond electronic health records (EHR)
- **Advantages**: improved diagnosis accuracy using word embeddings, similarity scores, and deep learning models, capability to consider multiple conditions instead of focusing on a single disease.
- **Disadvantages**: deep learning models in healthcare rely on datasets like MIMIC-III but inherit biases, limiting generalizability. Patient data is non-i.i.d., varying by region, making single-source models less reliable.
  - Needs high computation power.
- **Data accessibility**: MIMIC-III dataset which is publicly available. Semantic Corpus used is BioSentVec.
- **Code accessibility**: The code is accesible at code zip.

2. Paper 2:

- **Index-Title-Author**: 34 FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes by Tushaar Gangavarapu, Gokul S Krishnan, Sowmya Kamath, Jayakumara Jeganathan
- **Task**: Long-term disease prediction by leveraging unstructured clinical nursing notes, i.e. ICD-9 code group prediction task.
- **Innovation**: The innovation in this paper is that it uses free-text notes whereas prior models used structured EHR data. Uses Doc2Vec embeddings and non-negative matrix factorization for data transformation. Comprehensive deep learning benchmarking across multiple different architecture models. Improves performance AUPRC and AUROC performance over state of art (SOTA) models.
- **Advantages**: Takes into account nursing notes, model using nursing notes beats SOTA models based on EHRs. No dependence on structured EHR allows medical facilities to use these models in developing countries. As the infra to get an EHR in place may not be easy.
- **Disadvantages**: Dealing with noise, errors and inconsistencies, data preprocessing can get complex -> cleansing, toknization, etc
- **Data accessibility**: MIMIC-III dataset which is publicly available. NLP resources: Doc2Vec embeddings, Non-negative matrix factorization (NMF), NLTK, Stanford NLP and GENIA for preprocessing
- **Code accessibility**: No open source repository but implementation details are available in the paper.

3. Paper 3:

- **Index-Title-Author**: 35 SurfCon: Synonym Discovery on Privacy-Aware Clinical Data by Zhen Wang, Xiang Yue, Soheil Moosavinasab†, Yungui Huang, Simon Lin, Huan Sun

- **Task**: Synonym discovery using privacy-aware clinical data.
- **Innovation**:
  - Instead of relying on raw text, SurfCon tries to further abstract out things by operating on extracted medical terms and co-occurence counts which preserves patient privacy.
  - Captures both character and word level information.
  - Handles out of vocabulary (OOV) terms even if terms are not present in training data.
- **Advantages**: Privacy perservation which is a HUGE plus, effective synonym matching, handles noisy data.
- **Disadvantages**: Loss of local context since the framework uses co-occurence statistics and not the nursing notes, compute costs are high since the bi-level encoding and dynamic context matching require significant training time.
- **Data accessibility**: Data from Stanford Hospitals & Clinics Medical Term Co-Occurence dataset.
- **Code accessibility**: Available on Github.

4. Our target paper

- **Which paper will we replicate?** We will replicate the FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes, index 34.
- **Scope of replication (IMPORTANT)**: We intend to use existing models and train a few of our models to save time. Especially since there are a lot of nursing notes and the model weights are not publicly available. For example, 5 deep learning models for ICD9 prediction on MIMIC-III (1.4) dataset.
- **Task**: Long-term disease prediction by leveraging unstructured clinical nursing notes, i.e. ICD-9 code group prediction task.
- **Why choose this paper?** FarSight presents an innovative idea to approach disease prediction by using unstructured clinical text. The study also demonstrates significant improvements in disease prediction accuracy.
- **What are the specific hypotheses from the paper that you plan to verify in your reproduction study?** 1) The fact that unstructured notes will give better performance compared to structured EHR data. 2) FarSight's long-term aggregation can detect disease onset by leverarging historical notes. 3) We plan to further ENHANCE this by employing a guided and unguided Mixture of Expert models that can make use of the different architectures in the FarSight paper.
- **How are you assured that you can obtain appropriate data and computational resources?** The dataset used is MIMIC-III which ensures that we can reproduce the data. For compute, we plan to use Google Colab or rent a server on Runpod. Or possibly, even use Georgia Tech's PACE cluster.
- **What is the general problem in this work?** This work makes a good case that most model rely on structured electronical health records and tend to miss rich, patient-specific insights in unstructured clinical notes. This limitation causes model to show bad accuracy and delayed disease prediction. The FarSight model addresses this by leveraging unstructured clinical text improving performance and i.e. disease onset detection/prediction.
- **What innovations are in this work?** Already mentioned in section 2.
- **What advantages/disadvantages does the work have (e.g. accuracy to current problem is high/method is hard to be generalized)? What do you think could improve their method? Is their hypothesis legitimate?** Advantages/disadvantages already mentioned in section 2.
  - What could improve the model? Using hybrid data (both EHR and unstructured text), use newer model architectures such as LLMs and Transformers, employ a mixture of experts models. Add a qualitative aspect so that the people using this system can understand what the input text is the model lookng at.
  - Is the hypothesis legitimate? Yes, the hypothesis is well supported by results from the MIMIC-III dataset showing significant performance over existing SOTA models on structured EHR data.