The Kashmir Hub for Artificial Intelligence Research

(The KHAIR)

Master/Bachelor Internship Projects

Contact Information:

Website: `https://thekhair.github.io`

February 3, 2026

Project 1: Dementia Detection based on Multi-omics Integration

Project Description: Develop AI models that fuse multiple omics data into a unified diagnostic system for early, precise dementia detection.

Targets: This project will develop deep learning (DL) models that fuse multiple omics data types into a unified diagnostic system to enable early and precise detection of dementia. By integrating high-dimensional biological data—such as genomics, transcriptomics, proteomics, metabolomics, and clinical measures—AI can uncover subtle molecular signatures and complex interactions that precede overt cognitive decline, potentially transforming routine diagnostics and personalized prognosis.

Dementia, including Alzheimer's disease (AD) and related neurodegenerative disorders, has a multifactorial molecular basis involving genetic variants, protein dysregulation, metabolic changes, and altered cellular pathways. Single-modality biomarkers (e.g., imaging or cognitive scores) capture only a fragment of this complexity, limiting diagnostic sensitivity and specificity. In contrast, multi-omics integration combines orthogonal biological layers, enabling richer representations of disease mechanisms that improve diagnostic accuracy and subtype differentiation.

The foundational dataset for this project is **ANMerge**, a comprehensive Alzheimer's disease cohort accessible through the Synapse platform (Synapse: `syn22252881`). This resource includes longitudinal, patient-level, multimodal data, encompassing genetic profiles, proteomic and transcriptomic measurements, and detailed clinical annotations. Ethical access requires registration and data-use approval, ensuring research integrity and compliance.

Plan: You will start learning and understanding omics data including Genomics, Transcriptomics and so on and learn about Dementia/Alzheimer's and prepare a short literature survey on what has been done till now. Based on the prepared survey, we will investigate how we can further improve the models.

Hardware: GPU 24G.

Related Resources:

- Multimodal data (Synapse / ANMerge): `syn22252881`.
- Paper (PubMed): PMID `33285634`.

Project 2: Alzheimer's Disease Progression Modelling: Dementia Detection Via Speech

Project Description: Develop AI models that analyze speech patterns (acoustic, linguistic, semantic) to detect early-stage dementia through non-invasive voice recordings, transforming speech into a digital biomarker for cognitive health monitoring.

Targets: This project aims to develop AI models that detect early-stage dementia by analyzing

patterns in spontaneous speech recordings, transforming voice into a non-invasive digital biomarker for cognitive health monitoring. Speech changes — including acoustic markers (e.g., pauses, pitch, tempo), lexical and grammatical usage, and semantic coherence — have been shown to correlate with cognitive decline and can serve as early indicators of dementia long before clinical diagnosis. While ADReSSo offers a snapshot for classification, recent research has highlighted the importance of longitudinal multimodal datasets, which track participants over time across multiple communication modes — including spoken conversations, transcriptions, typed text, and extra-linguistic signals (e.g., pen and keyboard dynamics). These richer datasets make it possible to model cognitive changes over extended periods, providing deeper insights into disease progression beyond a binary classification task.

Plan: You will start investigating the state of the art models for paralinguistic speech such as wav2vec2, whisper, etc. and how we can integrate longitudinal data in monitoring the progression of Alzheimer's disease. Hardware: GPU 24G

Related Resources:

- Paper: https://arxiv.org/pdf/2502.19208v1
- Longitudinal Multimodal Data : https://link.springer.com/article/10.1007/s10579-023-09718-4
- Data: https://talkbank.org/dementia/ADReSSo-2021/index.html
- Data Source: ADReSSo Dataset: The ADReSSo (Alzheimer's Dementia Recognition through Spontaneous Speech only) dataset provides a standardized benchmark for this task. It contains balanced spontaneous speech recordings from cognitively normal individuals and people with Alzheimer's, collected via controlled picture description tasks. It encourages models that work directly from speech without relying on manual transcripts, although automatic transcription is also used.

Project 3: Drug Target Discovery Using GNNs and Knowledge Graphs

Project Description: Drug–target discovery plays a crucial role in accelerating the drug development process, yet it is often hindered by the complexity of interactions within biological systems. This project aims to tackle the drug–target prediction problem using advanced models like GNNs and MuCoS framework. We conceptualize drug–target prediction as a link prediction task within heterogeneous biomedical knowledge graphs (KGs) that incorporate various entities, such as drugs, proteins, diseases, and pathways, creating a comprehensive visual representation of their interactions. Traditional knowledge graph embedding techniques, such as TransE and ComplEx, fall short due to their dependence on computationally intensive negative sampling methods and limited generalization capabilities to unseen drug–target pairs. The MuCoS model addresses these limitations by leveraging high-density neighbor sampling, effectively identifying and capturing essential structural features while integrating contextual embeddings from BERT. This project will involve creating a new dataset for link prediction tasks that extends current KGs with anatomy info as well as develop new MuCoS and GNN-based methods to be applied to the new and other existing datasets like KEGG50k and PharmKG-8k.

Targets: The end goals of the project will be a new dataset, a new model and a publication.

Plan: You will start with understanding BERT and GNNs and how to use these models for prediction; learning about knowledge graphs and link prediction; learning about drug target interaction datasets such as kegg above and implementing MuCoS to reproduce the results; design a new dataset, test MuCoS and GNN models on it; Plan a new model based on the new dataset structure and features and validate its superiority; write up your work.

Related Resources:

- Hardware: GPU rtx 3090 with 12GB vram
- Data: datasets such as PharmKG info to find disease–gene–drug links
- Paper: https://aclanthology.org/2025.bionlp-1.27/

Project 4: Investigation of Tokenization for Kashmiri Language

Project Description: This internship project will systematically investigate and benchmark tokenization methods (such as byte pair encoding) for diacritic-rich, low-resource languages, with Kashmiri as a primary case study. Most existing tokenizers are optimized for English and other non-diacritic languages, leading to substantial performance degradation when applied to languages like Kashmiri that exhibit extensive diacritical usage and complex morphological structures. This research will analyze how these linguistic characteristics impact tokenizer behavior and evaluate their downstream effects on key NLP tasks, particularly Kashmiri–English neural machine translation. Core Challenge: Kashmiri represents a particularly challenging scenario with dual resource constraints: limited digital corpora and a complete absence of standardized tokenization approaches designed for its diacritic-intensive Perso-Arabic script. This creates a critical bottleneck in developing functional NLP tools for Kashmiri's 7+ million speakers.

Targets: The end goal of this project is to build a tokenizer for Kashmiri language and a publication/scientific report and to build an embedding model.

Plan: You will start understanding Kashmiri language morphology structure, and learn about tokenizers for other languages such as English, Persian and Arabic and adapt those tokenizers for Kashmiri Language.

Related Resources:

- Hardware: GPU 24G
- Dataset: To be added soon