Deep Learning-based Imputation Techniques for Handling Missing Values in Healthcare Data

Supervisor: Dr. Farnaz Rahimi

Electronic Health Records (EHRs) often contain missing values due to various reasons, such as data entry errors, inconsistent documentation, or differences in clinical workflows. Missing data can negatively impact the quality of analyses and predictive models in healthcare, especially when working with sensitive patient data. Accurate imputation of these missing values is crucial for improving data quality, enabling more reliable predictions, and enhancing downstream predictive modeling tasks.

This project aims to explore and compare state-of-the-art machine learning techniques, with a particular focus on masked autoencoders and variational autoencoder-based networks, for imputing missing laboratory test values in the publicly available MIMIC-IV dataset. The performance of these imputation methods will be assessed to evaluate their effectiveness and impact on subsequent predictive tasks.

The project will proceed in the following steps:

- Investigate various machine learning and deep learning methods for imputing missing values in EHR data.
- Implement and evaluate masked autoencoders and variational autoencoder-based models for missing data imputation.
- Compare the performance of deep learning-based imputation methods against traditional approaches, including tree-based models.
- Examine the impact of different imputation methods on downstream predictive modeling tasks, such as disease progression and mortality risk prediction.

Requirements:

- Proficiency in Python programming.
- Strong knowledge of machine learning models and their evaluation methods.