# 1 Goal

The overall goal of this project is to build a model to classify diagnoses in the ICU based on patient data. We used the MIMIC database, which contains data for over 40,000 ICU visits. The database is de-identified, and contains measurements for hundreds of variables, ranging from vital signs to lab results to chemical measurements.

# 2 Data

The paper uses a ”collection of anonymized clinical time series extracted from the EHR system at Children's Hostpical LA”, which differs in several aspects to the data in the MIMIC database. The most obvious of these is that the paper uses data from children, whereas the MIMIC database includes data from all age groups. Another potential difference is size; their data consisted of 10,401 pediatric intensive care unit (PICU) stays, and MIMIC contains at least four times that many ICU stays. However, after filtering out undesirable ICU stays, which can happen because there is not enough data or it is not in the population we are studying, we may end up with less than 40,000 ICU stays.

The MIMIC data is stored in a PostgreSQL database, from which we will pull the relevant variables for the ICU stays that we are interested in. For each variable we will create one or more summary variables, such as maximum, minimum, median absolute deviation, median, sum, number of times measured, and whether it was measured at all. During this project, we plan on imputing missing data based on what is clinically normal for a healthy patient.

# 3 Overview of Methods

We plan on using the methods used in the paper ”Learning to Diagnose with LSTM Recurrent Neural Networks”. This paper uses a Long Short-Term Memory (LSTM) network, a type of Recurrent Neural Network (RNN), to classify diagnoses, focusing on the 128 most common codes. It takes as input 13 variables, which are diastoic and systolic blood pressures, peripheral capillary refill rate, end-tidal $CO_2$, fraction of inspired $O_2$, Glascow coma scale, blood glucose, heart rate, pH, respirtory rate, blood oxygen saturation, body temperature, and urine output.

These variables are sampled at irregular times, so for every variable an hourly summary statistic was derived from the data. The paper uses the mean measurements, but it is possible to look at other types of summary statistics, such as maximum, minimum, variance, slope of the best fit line, etc. When a variable is completely missing, it is filled in with the clinically normal value. This is justified because if a variable is not recorded for some patient, it is probably because it appears normal; any abnormal measurements would have been taken note of.

The paper uses a target replication strategy to train the neural network, which gives an output at every sequence step instead of only at the final step. This outperformed RNN models that did not use target replication, so we will also try to apply this to our project. Target replication also helped in reducing overfitting. The paper describes several different methods that they used to combat overfitting and increase predictive power, which we will also try to apply.

## 4 Implementation

This project will be implemented primarily in python, using the Theano library to build the LSTM. Data will be pulled from the database with SQL queries, which we'll write to a .csv file and read into python to use as an input to the RNN.