Anthem

Rob Currie
Paula Alves
Robert Baertsch

Model Team
Harshitha Arul Murugan | Aman Prasad

Understanding Healthcare Data

Data Team
Cassidy Norfleet | Brendan Reilly-Langer

Baskin Engineering

Richard Jullig
Scott Davis
Golam Muktadir

Abstract

Compared to other wealthy countries, the U.S spends a disproportionate amount on **healthcare** with the **gap widening every year**.

We **utilized machine learning tools** to create models that predict a patient's future health conditions given their previous medical history.

Anthem, one of the largest healthcare companies in the U.S, is exploring the use of these models in order to provide **better care** for the patient and at a **reduced cost** for the care provider.

Key Terms

Synthea: Creates statistically accurate synthetic medical records curated by doctors for the purpose of research.

One-Hot Encoding (OHE): In a month, if a medical code was in a patient's history, the entry is 1 or 0 otherwise.

Docker: a platform to package up applications and code for the purpose of reliably running on different machines

Python libraries

Numpy: Contains numerical tools for computations.

Pandas: Helps organize data and perform calculations.

Pytorch: Creates, trains, and tests machine learning models.

Overview

The **goal** of our project is to use machine learning to **predict** the factors that lead to the **diagnosis** of a patient with **congestive heart failure. Anthem** can use these **predictions** to preemptively form a **treatment plan** and inhibit the progression of a patient's disease.

Machine Learning Terminology

Recurrent Neural Network (RNN)

Uses predictions from previous inputs to influence the current input's prediction.

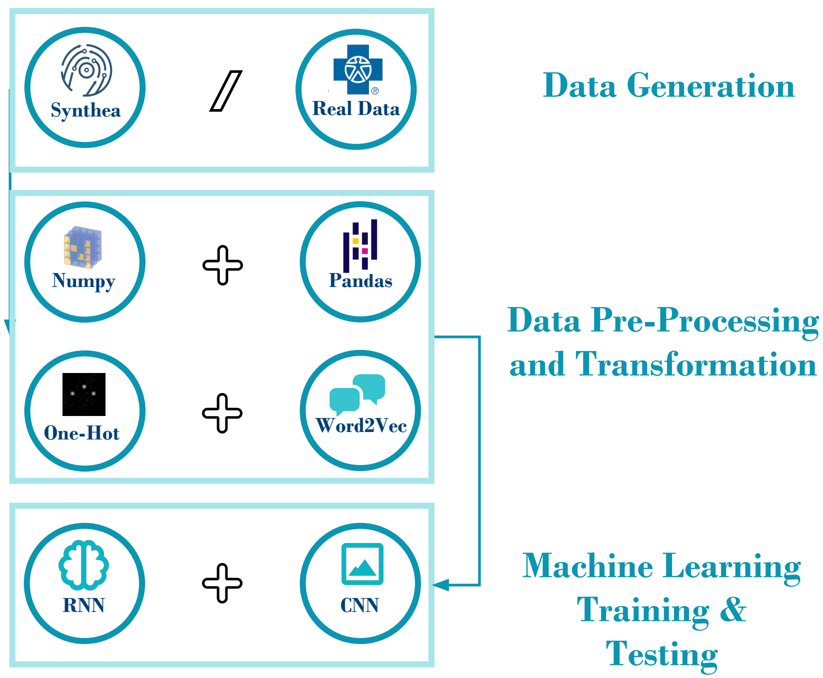
Convolutional Neural Network (CNN)

An image classification technique used to find unique patterns in different images.

Word Embeddings (Word2Vec)

Vectors of numerical values assigned to words or individual components in a given text.

Architecture



Approach

Because access to actual patient data is highly restricted, we used **Synthea** to generate artificial medical records. These records are used to create sparse **One-Hot Encoded** matrices **(OHE)** where each row represents one month of a patient's medical history and each column represents a unique medical code.

Using **OHE** matrices we trained a **CNN** and an **RNN**. We also created **word embeddings** for each of the medical codes present in a patient's history to train a **CNN**.

After developing successful models, we packaged our models in a **docker** container and sent it to **Anthem** to train on real anonymized patient data.

Results & Conclusion

For the **OHE** matrices built from the synthetic data, we were able to create a **CNN** that predicted outcomes at an accuracy of **96.36%**, and an **RNN** that predicted at an accuracy of **82.61%**.

In addition to OHE matrices, we also used the word embedding approach, which resulted in an accuracy of **79.34%** for the CNN and **71.53%** for the RNN.

Anthem then trained our models on the **word embeddings** processed from real data, which achieved an accuracy of **86.33%** on the CNN model and **83.72%** on RNN model.