Data Science Pipeline for End-to-end Advanced Clinical Decision Support System

Raine Hoang
k7hoang@ucsd.edu

Ojas Vashishtha
ovashishtha@ucsd.edu

Rohan Duvur
rduvur@ucsd.edu

Kyle Shannon
kshannon@ucsd.edu

Abstract

The rapid growth of electronic health records (EHRs) and the increasing availability of clinical data have created new opportunities for developing advanced clinical decision support systems (CDSS). These systems can assist healthcare providers in making quick, informed decisions by leveraging the power of data science and machine learning techniques. By using the MIMIC-IV and MIMIC-CXR databases, we aim to develop a comprehensive data science pipeline for an end-to-end advanced CDSS, focusing on the early detection and risk assessment of sepsis, a life-threatening condition. Our project utilizes a ResNet-50 CNN model to detect pneumonia from chest X-ray images. We combine the output of the ResNet model with patient data to build a CatBoost model that estimates the likelihood of developing sepsis.

Website: https://kshannon-ucsd.github.io/24wi-capstone-profile/
Code: https://github.com/kshannon-ucsd/24wi-dsc180-project

1 Introduction 2
2 Methods 3
3 Results 7
4 Discussion 8
5 Future Work 10
References 10

1 Introduction

Sepsis occurs when the body's immune system has an extreme reaction to a disease, resulting in major organ dysfunction. It is a fast-moving disease that can leave one dead in a few days, making time the key determinant in a patient's survival. However, the diagnosis of sepsis remains difficult and time-consuming due to the complex and heterogeneous nature of the condition.

Clinical decision support systems (CDSS) intend to assist medical professionals in their decision-making. The paper ”An overview of clinical decision support systems: benefits, risks, and strategies for success” (Sutton et al. 2020) informs us that these systems are connected to extensive databases that leverage a patient's information with the data it has to make recommendations on the next step. As a result, doctors can make faster and more informed decisions that could be essential to improving patient outcomes. The paper also identifies two main categories for CDSS: Knowledge-based systems and Non-knowledge-based systems. Knowledge-based systems relies on domain expertise as it has if-else statements encoded to make recommendations. On the other hand, non-knowledge-based systems use AI and machine learning techniques to figure out patterns within the data to make recommendations. We will be focusing on non-knowledge-based systems for early sepsis detection.

The development of a comprehensive data science pipeline for an advanced CDSS has the potential to revolutionize sepsis management and increase a patient's chance of survival. By providing healthcare workers with timely and accurate predictions of sepsis risk, the system can enable earlier interventions, targeted therapies, and personalized care. Moreover, the pipeline serves as a framework that can be extended to other clinical domains and decision support tasks, paving the way for more intelligent and data-driven healthcare systems.

1.1 Data

For this project, we used the MIMIC-IV (Johnson et al. 2020) and MIMIC-CXR (Johnson et al. 2019) databases provided by PhysioNet. MIMIC-IV is a relational database that contains deidentified medical information on over 200,000 patients admitted to the ICU or emergency department at Beth Israel Deaconess Medical Center. We also used the concept tables provided by the mimic-code GitHub to get some additional information on the patients such as their Elixhauser scores. MIMIC-CXR is built off of MIMIC-IV as it contains 377,100 chest X-rays stored as DICOM images that are all linked to a patient in MIMIC-IV through the patient ID and study times. Many people have taken the MIMIC-CXR database and preprocessed it in various ways such as segmenting the images and recording any metadata into concise CSV files. Additionally, they contained a CSV file that reported whether doctors found signs of pneumonia in the lungs. We obtained the metadata from the MIMIC-CXR-JPG database (Johnson et al. 2019) and the labels from Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification (Kermany 2018). For the first model, we used the chest X-rays along with the relevant metadata and pneumonia labels. Once we got the output of that model, we used it in combination with lab tests and vital signs to predict the development of sepsis.

2 Methods

2.1 Data for Pneumonia Detection

Our pipeline involved the development of two different models, meaning that we needed to acquire two appropriate datasets to train them on. The purpose of the first model was to identify pneumonia from a patient's X-ray. The inputs we needed for this model were images of chest X-rays and a label that indicated if the patient had pneumonia or not. We were able to obtain chest X-rays from the MIMIC-CXR database and the necessary metadata pertaining to those images from MIMIC-CXR-JPG.

We were able to use SQL and the Elixhauser indices given by the concept tables to query for these patients. We also got information related to when they stayed in the hospital, if they had sepsis, and the potential time they developed it. Once we had the SQL query, we loaded the data into Python to further merge it with additional CSVs provided by the additional databases mentioned above to match the X-rays to the correct hospital admission. To do so, we use the patient's subject_id to link the X-ray up to the proper patient and then used StudyTime to check if the X-ray was taken during their stay at the hospital. Finally, we merged with the Label Optical Co herence Tomography dataset to obtain the labels for pneumonia. The labels are encoded in the following way: 1 if there was pneumonia found and 0 if the lungs were normal.

2.2 ResNet-50 Model

2.2.1 Model Description

For pneumonia detection, we opted to extract features using the ResNet-50 CNN architecture. As our main input for this model was chest X-rays, we found CNN to be the best-suited model to perform our task due to its ability to discover underlying patterns in image data. We chose ResNet-50 for transfer learning due to its ability to generalize well to a wide variety of images and its robustness.

2.2.2 Model Architecture

ResNet-50 has the following architecture:

- **Input Layer**: Takes in 3-channel RGB images of 224 × 224.
- **Convolution Block**: Applies 64 filters with a 7 × 7 kernel, followed by max pooling.
- **Residual Learning Blocks**:
  - **Block 1**: 3 residual blocks, outputting 256 channels.
  - **Block 2**: 4 residual blocks, outputting 512 channels.

  - **Block 3**: 6 residual blocks, outputting 1024 channels.
  - **Block 4**: 3 residual blocks, outputting 2048 channels.
- **Classification**: Apply global average pooling followed by a fully connected layer for binary classification.
- **Output**: A sigmoid activation function predicts the probability of lung abnormality, where 0 indicates normal and 1 indicates abnormal.

2.2.3 Extended Model Layers

To further refine the feature extraction from ResNet-50 and enhance classification performance, we appended additional dense layers to the architecture. These layers introduce non-linearity, improve generalization, and aid in robust learning by leveraging batch normalization and dropout regularization. The additional layers are as follows:

- **Fully Connected Layer (Dense-512)**:
  - 512 neurons with L2 regularization ($\lambda = 5 \times 10^{-4}$) to prevent overfitting.
  - Batch Normalization to stabilize training.
  - ReLU activation function.
  - Dropout with a rate of 0.6 to improve generalization.
- **Fully Connected Layer (Dense-256)**:
  - 256 neurons with L2 regularization ($\lambda = 5 \times 10^{-4}$).
  - Batch Normalization.
  - ReLU activation function.
  - Dropout with a rate of 0.5.
- **Fully Connected Layer (Dense-128)**:
  - 128 neurons with L2 regularization ($\lambda = 5 \times 10^{-4}$).
  - ReLU activation function.
- **Output Layer**:
  - Single neuron with a sigmoid activation function to output the probability of pneumonia.

The final architecture follows a sequential design:

$$\begin{aligned}\text{Model} = \text{Sequential}([&\text{ResNet-50}, \text{Dense-512}, \text{BatchNorm}, \text{ReLU}, \text{Dropout},\\ &\text{Dense-256}, \text{BatchNorm}, \text{ReLU}, \text{Dropout}, \text{Dense-128}, \text{ReLU}, \text{Dense-1}, \text{Sigmoid}]).\end{aligned}$$

2.2.4 Model Implementation

The main tools we used to implement this model were Python, Tensorflow, and NumPy. For the optimizer and loss function, we used Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss. When it came time to train the model, we used a batch size of 32, set the maximum epochs to be 100, and enforced an early stopping patience of 5 epochs. The model is initialized with imagenet weights which are updates over time through training. In the early stages of this project, we opted to keep the imagenet weights fixed, but upon further testing we found that greater performance could be had by unfreezing the resnet layers.

2.2.5 Data Augmentation

Images had to be resized to 244x244 squares in order to be fed into the Resnet50 mdoel. To accomplish this, we first downsampled images to 256x256 squares then cropped from the center to get to the desired dimensions. To give our first model added rotation invariance and improve image clarity, we added slight rotations and brightness/ contrast boosts to our pre-processing pipeline. By doing this, we allow ResNet to avoid becoming dependent on specific hidden units.

We called a built-in Keras function preprocess input that takes image pixel arrays and normalizes them according to imagenet training data. The data is mean-centered and converted from RGB format to BGR

2.3 Data for Sepsis Risk Prediction

The second model took in the output of the first model along with some additional data to predict the probability of the patient developing sepsis. We followed UCSD Health's criterion for developing sepsis and used SQL to query for the appropriate features from MIMIC-IV. Additionally, we added a feature that indicated if a patient had pneumonia based off the ICD-9 codes.

2.4 CatBoost for Sepsis Risk Prediction

2.4.1 Model Description

For sepsis risk prediction, we used a CatBoost model to predict whether or not a patient will develop sepsis. While similar to a random forest model, we found that the catboost model performed far better on the negative (no-sepsis) class and had a higher AUC score as well.

2.4.2 Features

The following features were used for the second model:

- Heart rate
- Systolic blood pressure (SBP)
- Mean blood pressure (MBP)
- Respiratory rate
- Temperature
- Platelet count
- White blood cell count (WBC)

- Bands (immature white blood cells)
- Lactate
- International normalized ratio (INR)
- Partial thromboplastin time (PTT)
- Creatinine
- Bilirubin
- Pneumonia (binary)

The pneumonia feature is pulled from the results from the first ResNet model. These features were chosen due to both relevance to the target variable (sepsis) as well as the fact these features had the lowest number of null values.

To deal with null values for these features, we utilized probabilistic imputation, using the distribution of the features to impute meaningful values for the missing features, helping deal with our null values issue and expanding the number of meaningful features we could use.

2.4.3 Model Architecture and Implementation

We implemented the model using CatBoostClassifier from the catboost package as well as using a sklearn pipeline to combine standardization and prediction.

The following hyperparameters were used for the CatBoost model:

Table 1: Hyperparameters used for the CatBoost model.

Hyperparameter	**Value**
Iterations	3000
Learning rate	0.4
Depth	2
Min data in leaf	4
Subsample	0.8
Colsample by level	0.9
L2 leaf regularization	0.3
Random seed	42
Loss function	Logloss
Evaluation metric	AUC
Bootstrap type	Bernoulli
Verbose	False
Early stopping rounds	20

After splitting the data into train and test, we were able to make predictions and evaluate the performance of our model.

3 Results

Both models were evaluated using the AUC-ROC curve, accuracy, and the precision, recall, and f1-score on both the positive and negative class. These metrics ensure that we correctly predict as many positive cases as we can without over-predicting them. Figures 1 and 2 relate to the performance of the ResNet-50 model, and figures 3 and 4 relate to the CatBoost model.

3.1 ResNet-50 Performance

	Precision	Recall	F1-Score
No Pneumonia	0.92	0.8	0.86
Pneumonia	0.89	0.96	0.92
Accuracy			0.9
AUC			0.95

Figure 1: ResNet-50 Performance

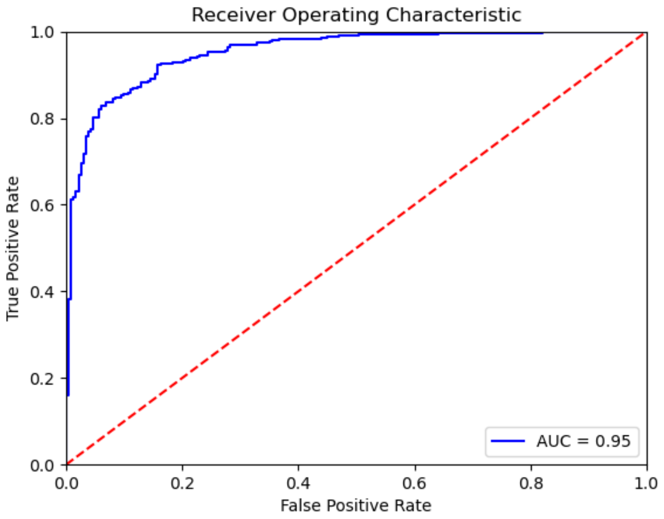

Figure 2: ResNet-50 AUC-ROC Curve

3.2 CatBoost Performance

	Precision	Recall	F1-Score
No Sepsis	0.78	0.86	0.82
Sepsis	0.79	0.69	0.74
Accuracy			0.78
AUC			0.86

Figure 3: CatBoost Performance

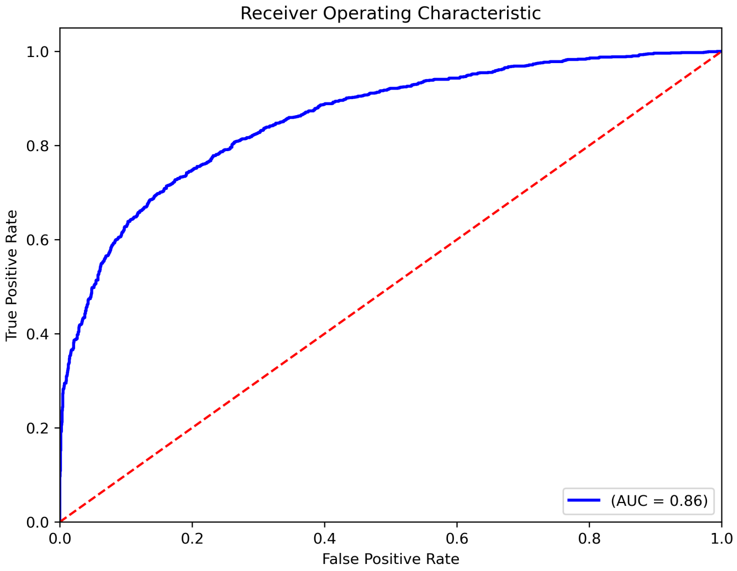

Figure 4: CatBoost AUC-ROC Curve

4 Discussion

4.1 ResNet-50

We can see that the ResNet model was able to detect pneumonia decently well and achieve a test accuracy of 90%. For the no pneumonia class, the high precision indicates that among all of the patients who were predicted to not have pneumonia, 92% of them didn't. The recall shows that we were able to detect 80% of the patients that did not have pneumonia. When it comes to the pneumonia class, 89% of those predicted to have pneumonia actually had pneumonia and 96% of patients with pneumonia were detected with the infection. Finally, the high AUC and the AUC-ROC curve shows that the model is able to differentiate between those with pneumonia and those without pneumonia well.

These metrics seem to suggest that the model is slightly over-predicting for pneumonia, but not by much. The 0.8 recall score for the no pneumonia class indicates that 20% of the people in this class are incorrectly diagnosed with pneumonia by our model. This is further supported by the lower precision for the pneumonia class as 11% of the people who were predicted to have pneumonia didn't have it. Those in this group could get unnecessary treatment, increasing costs for the hospital. However, this outcome is better than the alternative of the model not being able to detect someone with pneumonia, resulting in their death. As we can see, the model is able to predict 96% of pneumonia cases, meaning that it will rarely produce a false negative. Additionally, the high f1-scores for both classes indicate that the model is balanced and has strong performance.

4.2 CatBoost

The CatBoost model didn't perform as well as the previous model, but still does a decent job with a test accuracy of 78%. For the no sepsis class, 78% of those predicted to not have sepsis didn't have it and 86% of those who didn't have sepsis were correctly classified. When looking at the sepsis class, of those predicted to have sepsis, 79% of them did and of all patients with sepsis, 69% were correctly identified. Despite these lower scores, the model was able to obtain an AUC of 0.86, showing that it can distinguish well between the two classes.

Contrary to the ResNet-50 model, the CatBoost model seems to be under-predicting for sepsis. This is apparent in the recall score for the sepsis class as the model is only able to detect 69% of sepsis patients. Another way to think about this is that 31% of people with sepsis in the test set could not be correctly classified as having sepsis. Additionally, the sepsis class having a lower f1-score compared to no sepsis means that the model isn't able to detect sepsis as well as no sepsis. If this were to happen in a real clinical setting, this could lead to delayed treatment and possibly death. However, it is worth noting that while the sepsis f1-score is lower than the no-sepsis f1-score, both scores are still decently high, meaning that the model predictions are fairly balanced.

A reason as to why this model isn't performing as well as the previous could be due to the elusive nature of the disease. Sepsis is much harder to detect than pneumonia as sepsis shares many of its symptoms with other diseases. Since the diagnosis criterion for pneumonia is clearer than sepsis in the medical world, that will be reflected in how our models perform in comparison to one another. Another reason could be due to the quality of the data. For the first model, we were able to ensure that every X-ray used in the first model that no missing values. On the other hand, the data for the second model had a large amount of missing values that we had to impute probabilistically. Consequently, the values that we got may not have been the true value the patient had, leading to the second model not being as powerful as the first.

5 Future Work

Sepsis remains a leading cause of death worldwide, however, treatment of it is continuously improving. Our project shows a potential way to help decrease the time it takes to detect sepsis and improve a patient's survival chance by utilizing machine learning and CDSS. There is still much to be done as there are many ways to improve upon this pipeline. In the future, we would like to increase the performance of the CatBoost model to minimize false negatives as much as possible. To do so, we may use other medical databases besides MIMIC-IV and MIMIC-CXR to increase the amount of data available to train the model on. Additionally, we would expand this project out to other infections and organs. Sepsis can occur in response to any infection, so branching out of pneumonia and lungs could potentially save more lives. Finally, in the long term, we would use federated learning to ensure data privacy and security while also obtaining more training data.

References

Johnson, Alistair, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. "Mimic-iv." *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*: 49–55

Johnson, Alistair EW, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." *Scientific data* 6 (1), p. 317

Johnson, Alistair, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. "Mimic-cxr-jpg-chest radiographs with structured labels." *PhysioNet* 101: 215–220

Kermany, Daniel. 2018. "Labeled optical coherence tomography (oct) and chest x-ray images for classification." *Mendeley data*

Sutton, Reed T, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. "An overview of clinical decision support systems: benefits, risks, and strategies for success." *NPJ digital medicine* 3 (1), p. 17