FDS Project

Laksh Kataria (lvk8525) | Harsh Harwani (hh2752) | Anurag Mishra (aa9279)

1. Problem Statement:

It’s a widely accepted reality that there is a rampant substance abuse problem in our world today. It has become an essential task to devote resources and technology to help tackle this crisis. Data Science can be leveraged to study and find patterns that would help hospitals to understand patients, their needs, and if they can be discharged or not. Studying various features can help doctors predict the discharge status of a patient.

A previous study compared various machine learning models with Super learning and reported results based on comparing the models. Our project aims to apply predictive machine learning models to help hospitals identify the likely course that the treatment provided to a patient coming for substance abuse disorders will take, i.e, whether the treatment regime would be successful or unsuccessful. Our final dataset will have a target variable ‘Discharge_Status’ with class ‘1’ indicating successful treatment completion and discharge and class ‘0’ for an unsuccessful treatment program. Further, we will also be providing descriptive analysis on - state-wise death rates, most lethal addictions, and most susceptible age groups. Moreover, we will be using explainable AI tools such as LIME, and Shapely Values, to interpret our model results.

2. Background:

A study published in the journal Plos One, in April 2017, looked at the TEDS-D 2006–2011 data set. The TEDS-D dataset contains information for patients that are admitted to hospitals across the United States, specifically for substance abuse-related health issues. This data is collected by the Substance Abuse and Mental Health Services Administration, under the United States Department of Health and Human Services.

The article referred to above aims to introduce Super Learning, which is a Machine Learning ensemble method, and compare its performance against other predictive algorithms, commonly used in ML, such as deep neural networks, random forests, and logistic regression, to match the patients, based on their characteristics encapsulated in the TEDS-D dataset, to the appropriate SUD (substance use disorder) treatment that they must receive. This paper forms the inspiration for our work, where we aim to deploy predictive ML algorithms, to predict the discharge status of incoming patients and identify at-risk groups and the most lethal addictive substances, amongst other exploratory analyses.

3. Dataset Description:

As alluded to earlier, we will be using the TEDS-D dataset, collected by the Substance Abuse and Mental Health Services Administration, for the year 2019. The data set in its entirety contains 1.7 million rows and 76 columns, where each row corresponds to the data collected for a single patient, and each column holds information for that patient. Each of the features in the data set is categorical and has been encoded to display numerical values corresponding to their class. The feature columns are also nominal.

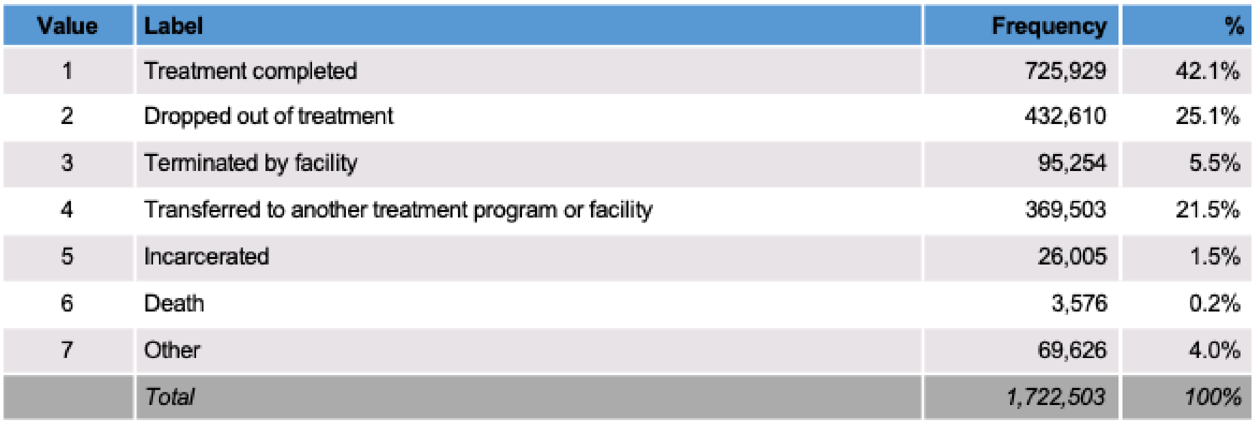
3.1 Data Reduction Steps:

Feature Selection -

The data set contains a total of 76 features. We apply pre-analysis feature selection techniques such as correlation analysis, and domain knowledge to drop 27 features, thereby ending up with a data set with 1.7million rows and 49 features

Downsampling -

Figure 1 shows the description of our target variable, as included in the original dataset. The reasons for handing a discharge to a patient are specified and encoded, as shown in the table. In our project, we narrow down the objective to a binary classification, wherein we predict if the treatment provided to the patient was completed, i.e, a ‘Successful Discharge’ or incomplete, i.e, an ‘Unsuccessful Discharge’



Value	Label	Frequency	%
1	Treatment completed	725,929	42.1%
2	Dropped out of treatment	432,610	25.1%
3	Terminated by facility	95,254	5.5%
4	Transferred to another treatment program or facility	369,503	21.5%
5	Incarcerated	26,005	1.5%
6	Death	3,576	0.2%
7	Other	69,626	4.0%
	Total	*1,722,503*	*100%*

Fig 1: Dataset description - Initial

- To reduce the number of data points within our dataset, we start by dropping all rows having target values - 2,6,7 (see assumptions for reasoning)
- To convert the problem to a binary classification, we encode classes - 3,4,5 as class ‘0’, and class ‘1’ remains the same
- We now end up with a dataset having 1.2 million rows and 49 features
- Now, we downsample our dataset further, making it easier to process and store:
    - We drop all rows containing any ‘Null’ values. This approach leaves us with a dataset having 71k rows and 49 features, with a class split of 41k and 30k for classes ‘0’ and ‘1’ respectively

The entire dataset has been documented in this pdf, and a description of each of the features included in the dataset can be obtained from the same.

Exploratory Data Analysis:

Before we set up a model pipeline and perform predictions on the dataset, we carry out some basic Exploratory Data Analysis.

We perform our analysis in two ways -

1. Exploratory Analysis on the entire dataset: For this analysis, we use the downsampled dataset. Through this analysis we identify:
    - The most common substance addictions per Race, for example - In our dataset, it was identified that category ‘White’ were more prone to Heroin addiction, and ‘American-Indian’ & ‘Asians’ were more susceptible to alcoholism.

    - The most common substance addictions per age group. It was identified that people in the age range ‘25-29’ were the primary Heroin abusers, followed by the age group ‘30-34’. Similarly, most Methamphetamine abusers also belonged to the age group ‘30-34’.

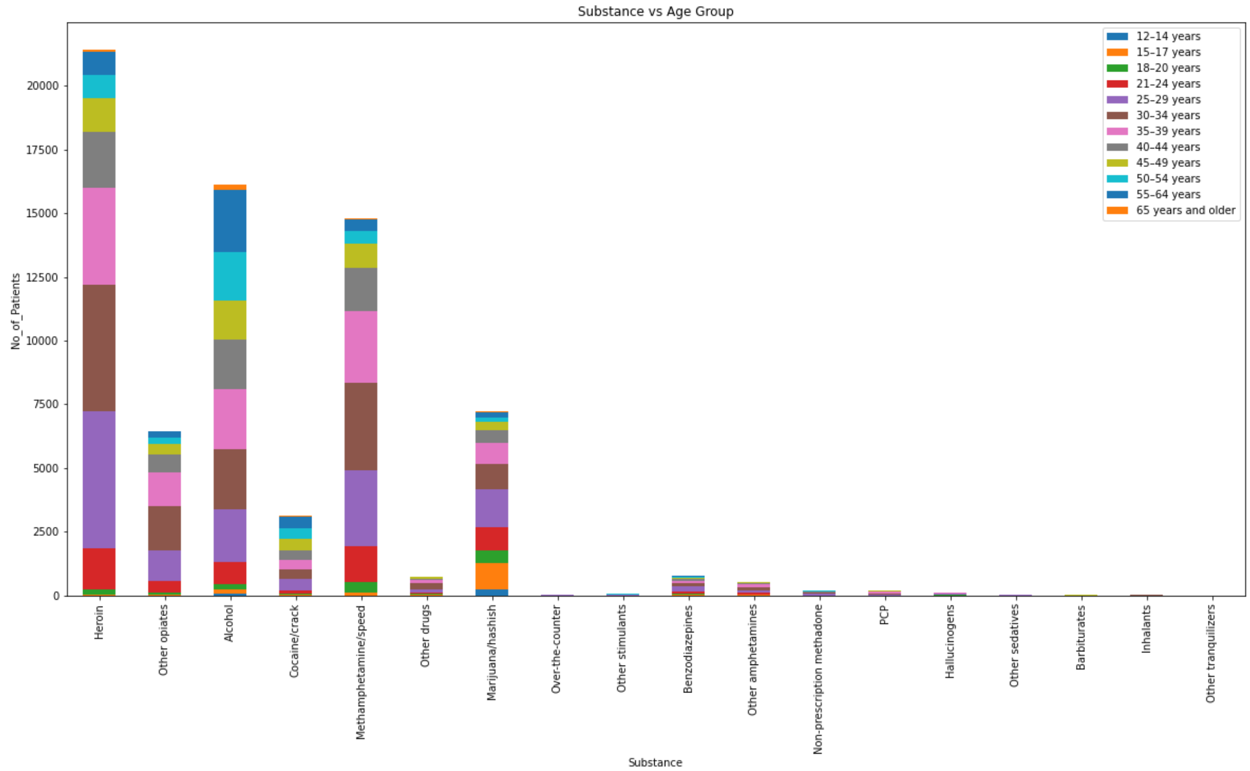


Fig: Chart showing most common age group per substance

Further, we also performed analysis to identify the ‘Length of Stay’ within the medical facility, and visualized the treatment outcomes per age group.

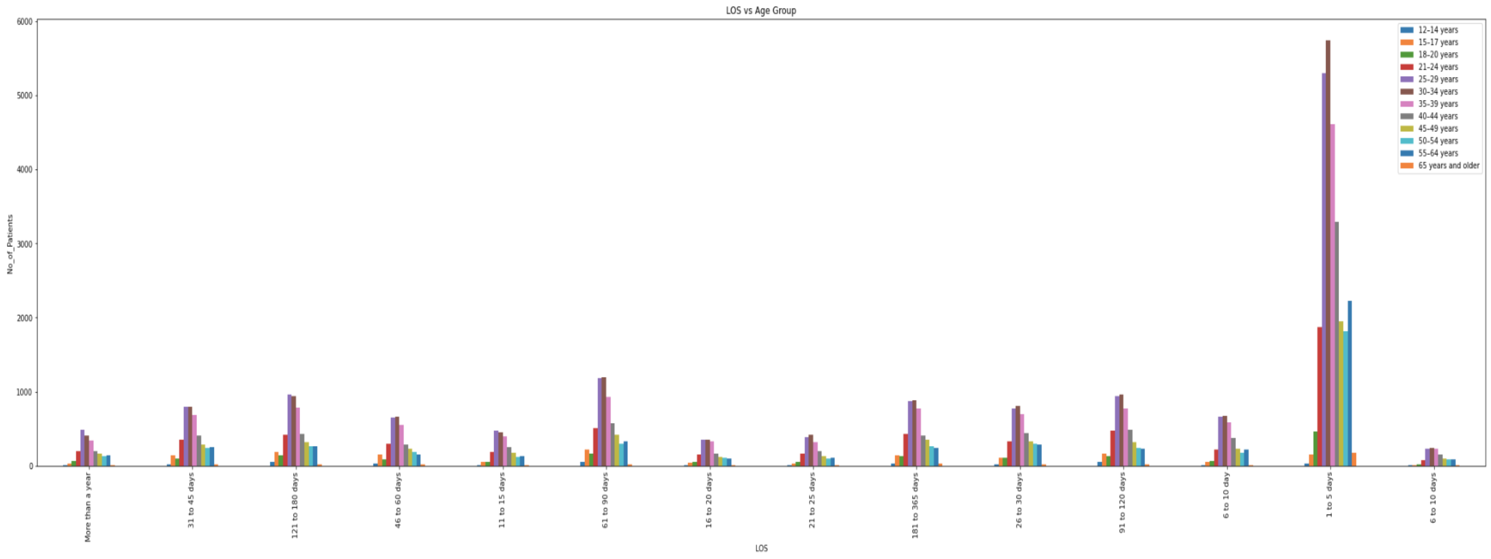


Fig: Plot showing ‘Length of Stay’ for each age group

2. Analysis on specific treatment outcome - ‘Death’: In our dataset, class ‘6’ within the column ‘Reason’ corresponds to a fatal outcome. Upon data analysis, we realized that approximately 3500 patients died within the treatment facility. We felt

the need to perform a special analysis on this specific dataset to identify any common patterns or addictions which could lead to this outcome.

- First, we try to find the most common substances that these patients were addicted to. Heroin addiction was found to be the most fatal and by quite some margin.

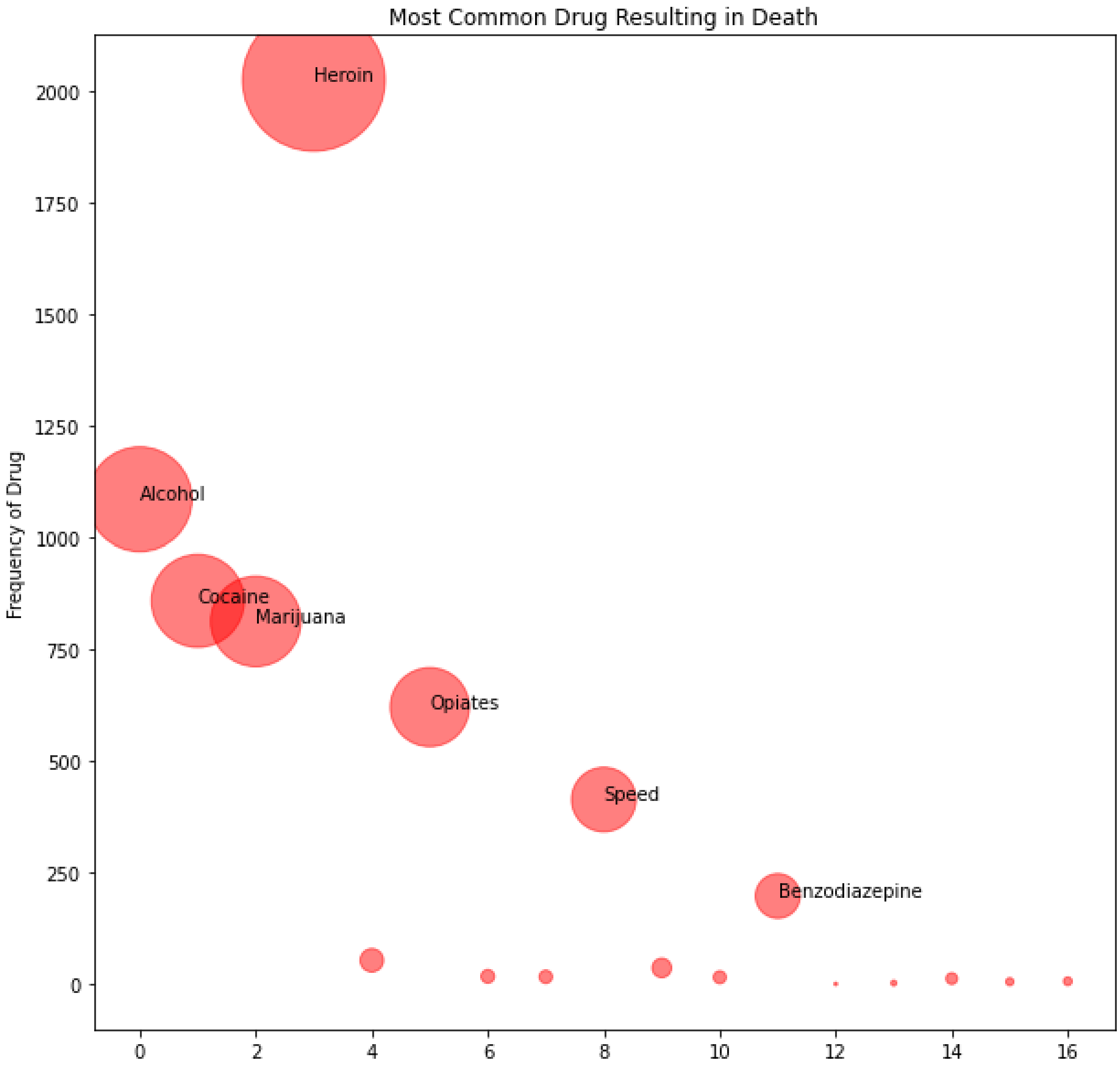


Fig: Bubble Plot showing most fatal drug addictions

- Second, we identify the age groups having most fatalities. Unsurprisingly, most deaths occured in the age group ‘55-64’, however, perhaps the most surprising result from the analysis was the fact that age groups ‘21-24’ and ‘65+’ had almost the same number of deaths.

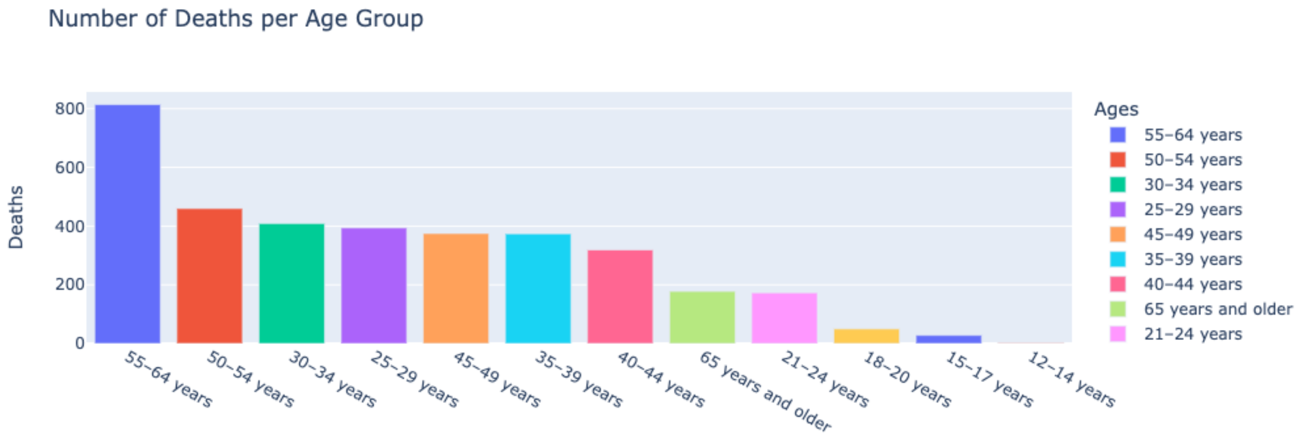


Fig: Chart showing death rates per age group

- Finally, we visualize the death rates across the various states of the USA and present our findings in a chloropleth map. California and New York have the highest death rates, with New York having almost double the count of California.

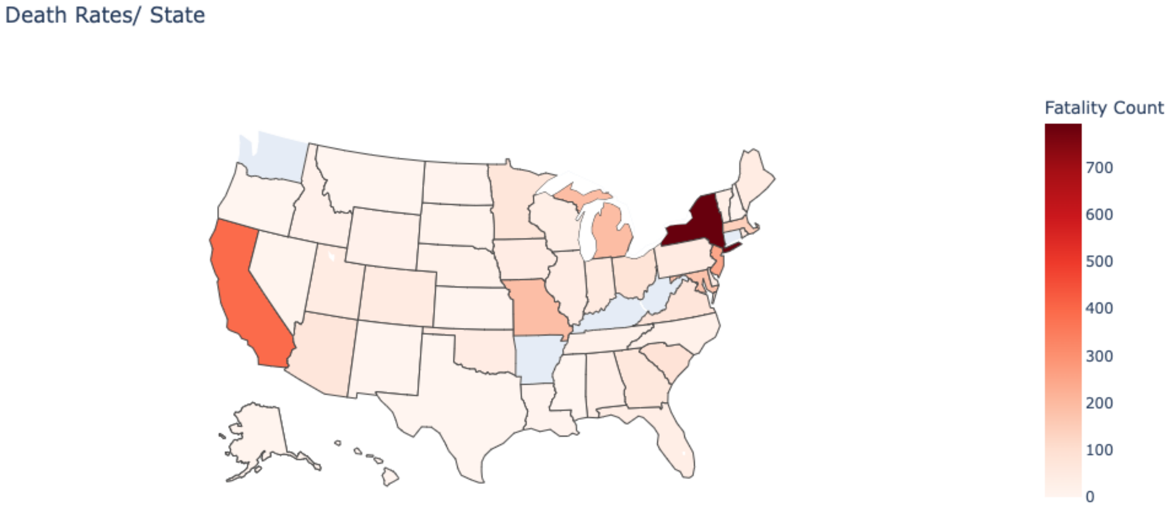


Fig: Chloropleth map depicting death rates across the USA

Model Type and Description:

As mentioned earlier, our primary goal is to predict the successful treatment of patients admitted to the healthcare facility and identify the key features which could help provide hospitals with prior knowledge at the time of patient intake, regarding the possible outcomes for that case. For this purpose, we will be building a predictive model.

Pipeline, models, gridsearchcv, hyperparameter tuning, lime shapely, scaling, best model

Evaluation metric :

The evaluation metric that we will be using for our model, would be the **Precision** of the model. For our model, reducing the number of 'False Positives' is more important as compared to 'False Negatives' (check assumptions). To measure the ability of our classifier, we will be using the ROC-AUC curve.

Results:

Based on the pipeline structure as explained above, Random Forest classifier was found to be the best model, in terms of the evaluation metrics that we aimed to maximize (Precision). Apart from achieving a precision value of 0.88, the model also achieved an accuracy of 87% which gave it the edge over XGB, which was the second best performing model. To visualize our results, we created an ROC-AUC curve for the Random Forest model [insert image] [ss of precision recall values]

Interpretation:

Next Steps:

PERFORMING SIMILAR ANALYSIS ON MORE RECENT AND COMPLETE DATA

USING DEEPER MODELS SUCH AS NEURAL NETWORKS TO HANDLE HIGH VOLUME DATA

OPTIMIZING ALGORITHM TO REDUCE COST ON PATIENT AND HOSPITAL END

LEARNING MORE ABOUT OVERFITTING AND HOW WE CAN DEAL WITH THE OVERFITTING IN OUR ANANLYSIS

Performing similar analysis on more recent and complete data

Using deeper models such as neural networks to handle high volume data

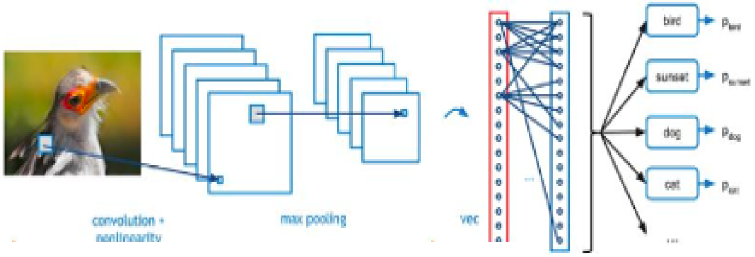
Optimizing algorithm to reduce cost on patient and hospital end

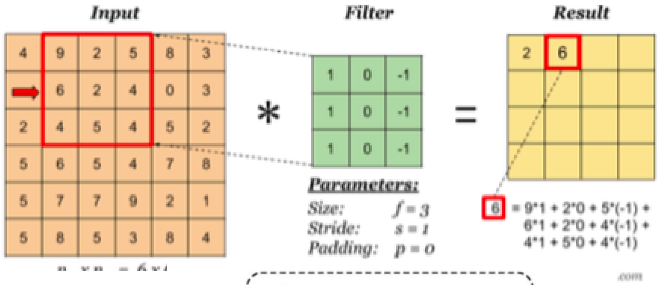
Learning more about overfitting and how we can deal with the overfitting in our ananlysis

Assumptions:
Since the dataset at hand comes from a government source, we assume the validity of the dataset is implied. We also assume that the risk factors and other important metrics which we identify through analyzing the information, albeit based on 2019 data, can be extrapolated to the current year and will provide a useful knowledge base for current world scenarios.

We dropped classes 2 - 'Dropped out of treatment', 6 - 'Death', and 7 - 'Other' from our dataset, as for classes 2 and 7, the treatment facility does not have information regarding the final status and whereabouts of the patient. For each of the included classes, the final status and condition of the patient are known. We drop class 6 as we perform a separate descriptive analysis on these patients, and rows corresponding to class 6 are only 3.5k which does not disturb the dataset.

Our reasoning for using Precision as our evaluation metric is that, since the model is being developed with a primary goal of helping hospitals identify if a specific treatment course would be successful or not, and not optimize the logistic costs of the hospital, falsely classifying a course as successful could have greater repercussions than a negative classification, in terms of the patient's health and well-being. Further, for an imbalanced dataset, precision would be the most appropriate choice for evaluation.



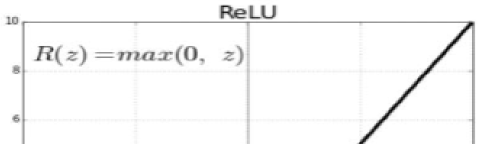


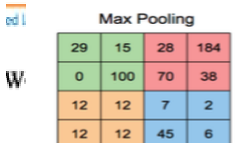
4	9	2	5	8	3
6	2	4	0	3	
2	4	5	4	5	2
5	6	5	4	7	8
5	7	7	9	2	1
5	8	5	3	8	4

1	0	-1
1	0	-1
1	0	-1

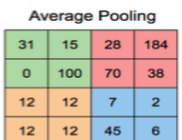
2	6		

6 = 9*1 + 2*0 + 5*(-1) +
6*1 + 2*0 + 4*(-1) +
4*1 + 5*0 + 4*(-1)

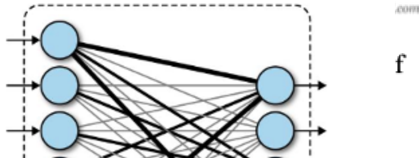


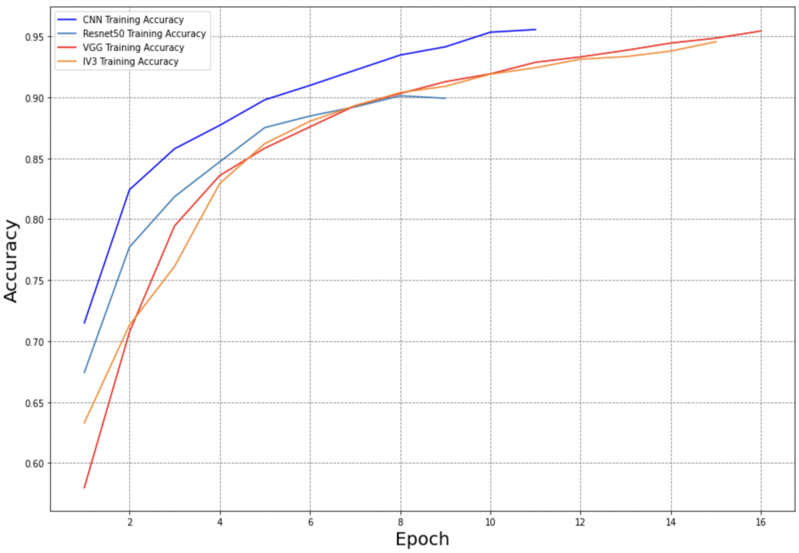


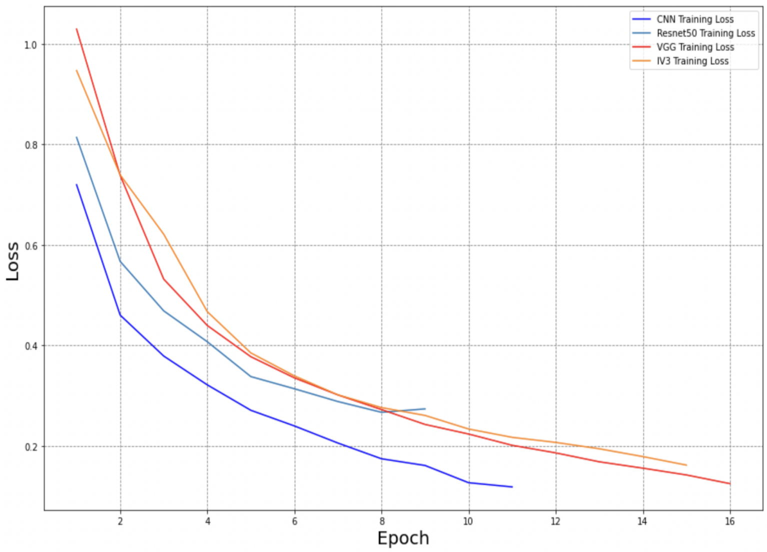
29	15	28	184
0	100	70	38
12	12	7	2
12	12	45	6

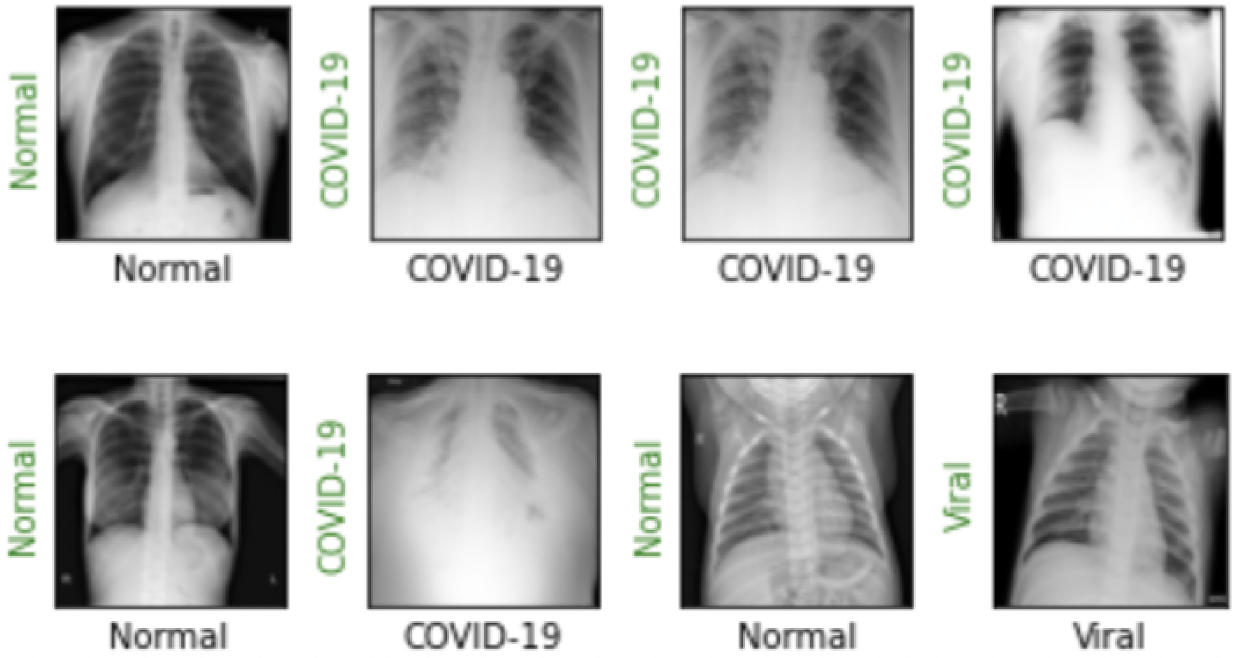


31	15	28	184
0	100	70	38
12	12	7	2
12	12	45	6




CNN Training Accuracy
Resnet50 Training Accuracy
VGG Training Accuracy
IV3 Training Accuracy
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
2
4
6
8
10
12
14
16
Epoch


CNN Training Loss
Resnet50 Training Loss
VGG Training Loss
IV3 Training Loss
Loss
1.0
0.8
0.6
0.4
0.2
2
4
6
8
10
12
14
16
Epoch


Normal
Normal
COVID-19
COVID-19
COVID-19
COVID-19
COVID-19
COVID-19
Normal
Normal
COVID-19
COVID-19
Normal
Normal
Viral
Viral