# 1. Introduction

## 1.1 Background

Healthcare sectors have seen significant improvements over the years by leveraging innovations in information and computer technology (Deloitte, 2019; U.S. Food and Drug Administration, 2016). Much of the gains have allowed more sophisticated methods in not only how information is stored and collected, such as that used in Electronic Healthcare Records (EHR) systems, but also increasingly played a role in how disease management and treatments are delivered. In the current digital world, this signifies the importance of healthcare data. It builds the foundation of creating and extracting knowledge from data to a piece of meaningful and possible novel information that can be enforced into clinical practices (Reddy, 2015).

Particular examples of practices in healthcare data and technology are data visualization and predictive modeling processes using the already established EHR database systems, which are particularly sought after for predicting mortality rates. Predictive modeling and analytics are a study on a set category of data aimed at making predictions about the future outcome based on historical data and specific statistical techniques and modeling (Chen et al., 2015). In the case of healthcare, predictive modeling can be used to find and detect risk factors for mortality in patients based on the patient's historical data, which is especially crucial in the case of intensive care units (ICUs) patients.

On average, there are four million ICU visits per year, with an average mortality rate of 8-19%, or 500,000 deaths annually (Mukhopadhyay et al., 2014). In addition to the mortality rates in ICU patients, critical care units also put a burden on the hospital institution's budget due to the length of stay and interventions done in ICUs (C et al., 1996). Due to these reasons, interest in predictive modeling and analysis has gained considerable interest in assessing ICU outcomes to

examine mortality risk and patients' length of stay in order to optimize resources and workforce energy expenditure for health institutions.

The development of innovative machine learning models and predictive analytical techniques such as decision tree-based random forest and boosting algorithms can be used to resolve the outlined challenges and facilitate the design of prediction models (Jones et al., 2008; Wolfe et al., 2005). By leveraging machine learning models combined with data mining techniques, it is possible to automatically extract valuable insights and find unique patterns in the patient's historical data. The topic involving machine learning for medical predictive analysis based on EHR isn't relatively new. Effectively, there has even been a previous study on the topic which has established a widely used benchmark that are used as a reference value for future researches (Harutyunyan et al., 2019).

Unfortunately, the results from machine learning predictive models are mostly aimed at researchers themselves instead of healthcare practitioners or decision-makers in healthcare institutions. Thus, more user-friendly products are needed to be effectively disseminated to report certain patterns in ICU data, which could lead to mortality or other factors that could contribute to better decision-making in the hospital institution itself. This visual problems have been addressed before in a study by R. Chen et al. (2015), where a web based visual platform was created to give a broader view of ICU based patient data. However, the study employs an outdated version of ICU data when a newer version has been released. In addressing the challenges mentioned before, this study proposed a visual predictive analytics platform that focuses on updated ICU patient database by leveraging machine learning models and interactive web-based platforms for result visualizations.

This study utilizes the Multiparameter Intelligent Monitoring in Intensive Care III ICU datasets (MIMIC-III), which is a freely available public database for critical care patients (Johnson et al., 2016; A. Johnson, Pollard, & Mark, 2016). The current dataset contains improvements from

the previous version of MIMIC-II. To date, there has been limited work that has been done in creating such interactive tools aimed for data exploration and prediction such as this. Though similar studies exist, such as that reported by R. Chen et al. (2015), the dataset used seemed to only consist of MIMIC-II and not MIMIC-III. For the current project, a machine learning model is used to create a clinical prediction task for the length of stay and mortality rates using a dashboard themed website. As a benchmark to measure the performance of the prediction tasks, the study aimed to exercise the benchmark data from Harutyunyan et al. (2019) as a reference value. Visualization of the data is done using Plotly Dash - a python framework used to create web applications - and will be connected to the MIMIC-III database along with the prediction model to provide a comprehensive view of the current clinical setting. Considering that the visualization framework mainly uses python, thus the machine learning model will follow suit by using available python packages to ease implementation with the web application. The proposed methods and implementation of the machine learning model towards predictive modeling and visualization will be described in the further section below.

## 1.2 Objective

The primary objective of this project will be to develop and evaluate machine learning models on a large set of electronic health record datasets of ICU patients. This project focuses on developing machine learning mode to explore prediction capabilities on mortality risk and length of stay of patients data from the corresponding datasets and to present the result using a visual platform to ease user understanding. This may help to gain knowledge and insights about the performance of the machine learning model towards the specific tasks along with a broader view of the underlying datasets used to develop the said model.

The objectives of this project are:

- To develop and test the best performing supervised machine learning model on predicting mortality risk and length of stay of ICU patients based on the proposed study benchmark.

- To develop a web dashboard application to provide a general view of the MIMIC-III datasets in a minimalistic and simple design following the main content and functionalities.
- To visualize and document the result of the model prediction tasks into the web dashboard application.

## 1.3 Significance, Scope and Definitions

The project uses EHR datasets obtained from the MIMIC-III database (Medical Information Mart for Intensive Care). MIMIC-III datasets are a large collection of unidentified patients datasets admitted to critical care units at large tertiary care hospitals in Boston, Massachusetts populations. The datasets are screened to obtain valuable and interpretable data using data mining methods further described in Chapter 3. The screening results are then subjected to machine learning and predictive analytical models to find possible models to predict mortality risk and length of stay of patients. Finally, the result will be summarized in a simple, interactive, and easy to perceive web application dashboard.