# MULTI-STAGE PROGNOSIS OF COVID-19 USING A CLINICAL EVENT-BASED STRATIFICATION OF DISEASE SEVERITY

BY

HAOTIAN CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Professor Ravishankar K. Iyer

# ABSTRACT

The COVID-19 disease has shown remarkable diversity in its manifestation. Precise anticipation of these manifestations is important to enable earlier intervention for high-risk patients and efficient deployment of medical resources. In this thesis, a multi-stage prognostic framework is developed for assessing COVID-19 patients at hospital admission and during disease progression. The analysis is conducted upon 10,123 COVID-19 patients treated at Rush University Medical Center at Chicago between 03/17/2020 and 08/07/2020. In order to characterize the patients with different severity, a stratification scheme is first established to assign patients to different stages of disease severity based on discrete clinical events (i.e., admission to hospital, admission to ICU, mechanical ventilation, and death). Then two prognostic frameworks were developed to predict the progression of COVID-19 through these stages: 1) a baseline model which uses the measurements collected at hospital admission to predict disease escalation to severe stages; 2) a progressive model which uses the measurements collected at the patient's latest stage to predict further escalation. It is found that future clinical stages can be predicted using baseline measurements with clinically significant accuracy. Finally, key risk factors are identified using Least Absolute Shrinkage and Selection Operator (LASSO) and decision tree algorithms. The developed multi-stage framework can be used to anticipate COVID-19 disease progression, allowing earlier interventions as well as better management of hospital resources.

# ACKNOWLEDGMENTS

First, I want to express my sincere gratitude to my adviser Professor Ravishankar K. Iyer for his extensive support and guidance during my graduate study. Professor Iyer is a very resourceful adviser, who provided me with the opportunity to complete my research on this frontline COVID-19 project. Moreover, he cared about my long-term success and always encouraged me to learn new knowledge.

Next, I would like to thank Dr. Yogatheesan Varatharajah for his help in maturing my methodology. I also want to thank Rush University Medical Center for making the high-quality data available for analysis.

Additionally, I am grateful to the members of the DEPEND research group, especially Yurui Cao, Chang Hu, and Kathleen Atchley, for their constructive feedback. The DEPEND research group provided a collaborative environment for me to conduct this study.

Finally, I want to acknowledge my family for their everlasting love and support.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ALB | Albumin |
| ALT | Alanine transaminase |
| ARDS | Acute Respiratory Distress Syndrome |
| AST | Aspartate Aminotransferase |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| BUN | Blood Urea Nitrogen |
| CRP | C-reactive Protein |
| DBIL | Direct Bilirubin |
| DIC | Disseminated Intravascular Coagulation |
| ICU | Intensive Care Unit |
| IMV | Invasive Mechanical Ventilation |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LDH | Lactate Dehydrogenase |
| NIV | Non-invasive Ventilation |
| NPV | Negative Predictive Value |
| PPV | Positive Predictive Value |
| RBC | Red Blood Count |
| RDW | Red Cell Distribution |
| $SpO_2$ | Oxygen Saturation |
| WBC | White Blood Count |

# CHAPTER 1

# INTRODUCTION

Since the emergence of COVID-19, first identified in Wuhan, China [1], a global pandemic has ensued [2]. Sustained transmission has been observed worldwide. As of July 17th, 2021, the virus has infected 84.5 million people and caused over 1.83 million deaths worldwide, including over 20.4 million cases and over 350,000 deaths in the United States.

Studies have been published on the clinical characteristics and treatment outcomes of COVID-19 in Chinese cities such as Wuhan [3], Shanghai [4], and Chongqing [5], as well as in New York City [6], [7]. These studies suggest that acute respiratory distress syndrome (ARDS) is a major driver of elevated mortality rates in critically ill patients. Several potentially impactful interventions have been found to reduce the severity of the illness and improve outcomes among this cohort: 1) early prone positioning (before intubation), which improves oxygenation and reduces the need for mechanical ventilation [8]–[10]; 2) Remdesivir therapy, which, when given to those requiring supplemental oxygen, can reduce recovery time [11]; and 3) dexamethasone, a corticosteroid and frequent adjunctive therapy for ARDS and sepsis, which has been found to reduce mortality among those with COVID-19 and respiratory compromise [12]. These findings suggest the importance of early identification of severe disease progressions, such as ARDS, as part of an overall strategy for treating COVID-19. Therefore, there is a critical need for a tool that can identify patients who should receive earlier interventions.

In this single-center study conducted at the highest-volume clinical center for COVID-19 in the state of Illinois, we characterized the evolution of infections, risk factors for infection, and predictors of severe illness (i.e., ARDS) in a multi-stage perspective. A multi-stage prognostic framework is then developed for identifying those at high risk. The framework consists of 1) triage models based on baseline laboratory measurements, vital signs, and demographics collected at hospital admission to predict the escalation to ICU admission, ventilation, and mortality, respectively; 2) progressive triage model which uses the measurements collected at the patient's current stage to predict further escalation (e.g., for the patients admitted into ICU, the model predicts the escalation to ventilation or mortality stages).

Several other previously published modeling approaches have attempted to evaluate predictors for clinical deterioration, mechanical ventilation, and death. Factors associated with disease progression include lower platelet and lymphocyte counts; increased markers of DIC, such as fibrinogen and d-dimer; increased LDH, AST, and CK, and abnormal CT scans [13]; clinical comorbidities, CRP, respiratory rate, and LDH [14], [15]; and higher SOFA score, age, and d-dimer levels [16]. One non-peer-reviewed study found similar factors and developed a nomogram for prediction [17]. In a peer-reviewed study, clinical comorbidities were the most predictive for severe disease progression [18]. However, the hospitals in different regions have different underlying populations, and thus the specificity of patients needs to be considered in the risk factor extraction. Our work has identified three critical indicators of the deterioration of COVID-19 patients, namely low albumin level, diminished $SpO_2$, and elevated white blood count, which are readily measurable in a clinical setting.

Although several studies have looked at COVID-19 disease progression, our study is unique as it involves a large urban population in a midwestern city in the United States and the

proposed multi-stage framework features a much finer granularity in terms of disease progression scenarios. The large patient sample in this study features diverse racial and ethnic groups in the Chicago area. Compared with New York and Los Angeles, the Chicago population has the largest percentage of African Americans and the youngest average age [19], [20]. Previous studies in other urban areas of the United States (i.e., New York and Los Angeles) have found strong differences in severe illness onset and outcomes across racial and ethnic groups [21]–[30]. These differences encourage a customized model to best characterize the patients in a certain area. In this context, our research characterizes the unique patient population in Chicago, Illinois.

Despite the distribution of the COVID-19 vaccine, it is still crucial to study COVID-19 disease since the vaccination coverage is far less than adequate. As of July 16$^{th}$, 2021, only 25.9% of the world population and only 1% of the people in low-income countries have been vaccinated [31]. Furthermore, even those vaccinated people can still be infected by COVID-19 [32]. Such an infection after vaccination, clinically called a breakthrough infection, is observed more often in the cases caused by the Delta variant. It was reported that the Pfizer-BioNTech and Oxford-AstraZeneca were only 33% effective against the Delta variant three weeks after the first dose [33]. Due to imperfect vaccine protection, those vaccine receivers are still at risk of COVID-19 infection. According to a survey conducted by *Nature*, almost 90% of scientist respondents expect the SARS-CoV-2 virus to become endemic [34]. In their vision, people will be facing COVID-19 disease for a long time and therefore the prediction for COVID-19 disease progression will continue being crucial.

The overall flow of our analysis is illustrated in Figure 1. Demographics and symptoms are presented for a total of 10,123 COVID-19 patients. Then a clinical event-based stratification

is provided, based on which the temporal trend of lab measurements, comorbidities, and radiological findings are presented in a multi-stage perspective. To identify the patients who are likely to enter severe stages, both initial triaging and progressive triage models are developed using logistic regression. Finally, key risk factors are identified using LASSO regression and decision tree algorithms. Compared with deep-learning models, the LASSO regression and decision tree are more data-efficient and can converge on a handful of patient records. In addition, these models offer higher interpretability to clinicians, which makes them more likely to be adopted in clinical practice [35].

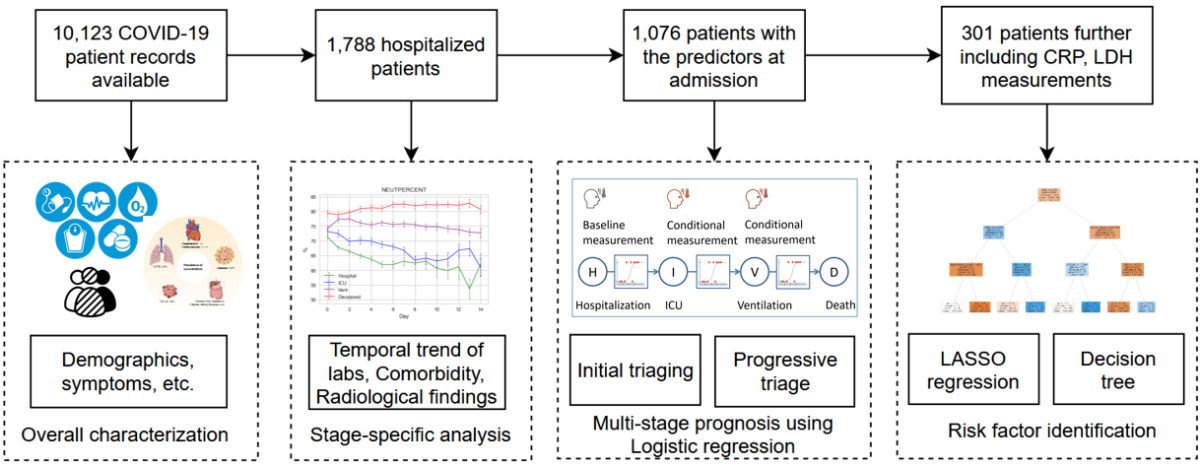



**Figure 1 Overarching framework of analyzing COVID-19 patients treated by a large urban hospital at Chicago.**

# CHAPTER 2
# BACKGROUND AND RELATED WORK

Diverse disease manifestations have been observed among the COVID-19 patients, progressing from asymptomatic presentation to critical illness such as severe acute respiratory syndrome and organ failures [36]. It is of paramount importance to characterize and predict the progression of COVID-19, which not only reveals the disease dynamics but also facilitates earlier intervention for high-risk patients.

In this chapter, we will first present background knowledge on COVID-19 disease progression. Then we will review prevalent models for predicting the disease progression. To systematically review them, these models are categorized into three types: 1) nomogram, a pictorial representation of mathematical formulas commonly used in current clinical practice; 2) traditional statistical learning models such as SVM, decision tree, random forest, and XGBoost; and 3) deep-learning approaches such as convolutional neural network (CNN) and recurrent neural networks (RNN). These methods provide diverse toolsets to predict COVID-19 disease progression, and some of them achieve high prediction accuracy. However, these methods only predict a single outcome (e.g., mortality), as shown in Table 1. In contrast, our multi-stage predictive model is able to output different levels of severity in terms of ICU admission, ventilation, and mortality. Built upon logistic regression, this predictive model is readily interpretable: the trained coefficients of logistic regression reflect the contribution of each factor to the outcome. Specifically, the signs of coefficients denote the positive or negative effects on the outcome, while the magnitudes of coefficients quantify the strength of these effects. Finally,

the model is developed using patient data in a large Chicago urban hospital containing diverse race and ethnicity groups.

## 2.1 Progression of COVID-19 Disease

The study of COVID-19 disease progression has attracted great interest due to its clinical value. It is found that the progression of COVID-19 disease involves multiple stages. For example, after studying a wide spectrum of clinical features including lab measures, vital signs, symptoms, and radiological findings, Siddiqi and Mehra [37] characterized the COVID-19 progression with the following phases:

1. Early Infection: This stage features mild constitutional symptoms. Lymphopenia can be observed in patients' lab findings but without other significant abnormalities. Treatment at this stage should mainly focus on symptom relief.
2. Pulmonary Involvement with and without Hypoxia: Patients in this stage typically have pulmonary disease, viral multiplication, and localized inflammation in their lungs. Radiological imaging may show bilateral infiltrates or ground-glass opacities in their lungs.
3. Systemic Hyperinflammation: This stage features the occurrence of extra-pulmonary systemic hyper-inflammation, which is also called a cytokine storm. Elevation of IL-2, IL6, CRP, Ferritin, and D-dimer can be observed among patients at this stage.

Because the above three-phase staging schema requires a substantial collection of clinical measurements which are not always accessible, an alternative staging approach is to monitor the transitions of patients in the hospital (e.g., the transition from ICU to ventilation) and directly use these clinical events for staging. Mody et al. stratified the patients based on the clinical units they

stay at, including 1) emergency department, 2) inpatient floor, 3) intensive care unit (ICU), 4) invasive mechanical ventilation (IMV), and 5) non-invasive ventilation (NIV) [38]. Then, they presented the statistics of patients among these stages. Their study showed that older male patients were more likely to be admitted to ICU, NIV, and IMV. They also found that compared to the patients with mild outcomes, those eventually intubated patients have more abnormal laboratory measurements at their baseline. Although no prediction model is proposed in this study, their findings indicate the predictive values of baseline features in anticipating disease progression.

## 2.2 Existing Models to Predict COVID-19 Disease Progression

With the COVID-19 disease progression characterized, clinicians and researchers would want to further predict which patients will develop severe progression. In the clinical setting, the identification of high-risk patients helps clinicians to deploy medical interventions.

Various models have been proposed to predict COVID-19 disease progression. They take subsets of clinical features as model input, including 1) demographic characteristics, namely age, sex, and race; 2) exposure history; 3) symptoms; 4) comorbidities; 5) laboratory findings; 6) vital signs; 7) radiological findings (including CT images); and 8) treatments. Most models have mortality as their output, others focus on ICU admission or a certain severe stage they define. These methods can be categorized into three types, and their details are presented in Table 1.

**Table 1 Existing models to predict COVID-19 disease progression**

| Type | Model | Output Label | Cohort Size | Features Used | Refe-rence |
|---|---|---|---|---|---|
| Nomogram | Parallel-scale nomogram | Severe outcome | 372 | Age, laboratory measurements (including LDH, CRP, RDW, DBIL, BUN, and ALB) | [39] |
| | | Mortality | 709 | Age, dyspnea, $SpO_2$, HCT, CRP, AST, and Ferritin | [40] |
| | | ICU admission | 1087 | Age, respiratory rate, systolic blood pressure, smoking status, fever, and chronic kidney disease | [41] |
| Traditional statistical learning | SVM, KNN, decision tree, random forest, and logistic regression | Mortality | 53 | Age, temperature, exposure history, clinical symptoms, laboratory findings, comorbidity, hospitalization, and treatment | [42] |
| | LASSO regression and logistic regression | Severe outcome | 1590 | Age, sex, comorbidities, laboratory measurements (neutrophil-to-lymphocyte ratio, lactate dehydrogenase, direct bilirubin), and chest radiography | [43] |
| | Multivariable logistic regression | Mortality | 299 | Age, lymphocyte count, lactate dehydrogenase and $SpO_2$ | [44] |
| | Random forest boosted by AdaBoost algorithm | Mortality | NA | Geographical location, travel history, symptoms, and demographics | [45] |
| | XGBoost | Mortality | 485 | LDH, lymphocyte, and high-sensitivity C-reactive protein | [46] |
| Deep learning | CNN | Mortality | 366 | CT images, sex, age, severity grade, and chronic disease | [47] |
| | RNN (with gated recurrent units) | Severe outcome | 2374 | Age, sex, comorbidities, and other unspecified EHR records | [48] |
| | CNN and RNN | Severity of lung pathologies | 42 | A series of X-ray images, collected throughout the hospitalization period | [49] |

### 2.2.1 Clinical nomograms

As is widely used in clinical practice, nomogram is a pictorial calculating instrument that applies a straightedge across the plot through the points on scales representing independent variables [50]. Then the straightedge crosses the corresponding datum point for dependent variables [51]. Those independent variables are the risk factors selected by experienced clinicians or from the multivariate risk-factor analysis.

The development of nomograms can be divided into two steps: 1) decide key risk factors and their coefficients; 2) project mathematical relationships into the diagram. Among 372 COVID-19 patients in Wuhan and Guangdong, Gong et al. found that more advanced age and higher levels of LDH, CRP, RDW, DBIL, BUN, and ALB on admission contribute to higher odds of severe COVID-19 [39]. These seven factors were then used to construct a nomogram, which resulted in AUC = 0.912 in the training cohort and AUC = 0.853 in the validation cohort. Acar et al. performed multivariable logistic regression on 709 patients and identified the higher age, comorbidity, dyspnea, low $SpO_2$, HCT, CRP, AST, and Ferritin as key risk factors for mortality, using which features a nomogram was constructed [40]. Similarly, Zhou et al. developed a nomogram based on age, respiratory rate, systolic blood pressure, smoking status, fever, and chronic kidney disease [41].

In summary, these nomograms visually represent the relationship between input risk factors and output risk score. One advantage of nomograms lies in their interpretability: the significance of each feature is reflected by its scale, and the relationship of these features is represented by a straightedge. However, nomogram methods have limited feature and function space: 1) while the risk factors for COVID-19 progression can be enormous, only a limited number of features can

be represented on a physical graph; 2) because the prediction equation has to be convertible to a straightedge in the nomogram, it cannot be arbitrarily complex.

### 2.2.2 Traditional statistical learning models

Statistical learning models overcome the limitation of nomogram methods by enlarging feature and function spaces. Jiang et al. experimented with prevalent machine learning methods including logistic regression, K-nearest neighbor (KNN), decision tree, random forest, and SVM in a COVID-19 patient cohort from China [52]. Their results showed the superiority of KNN and SVM for predicting ARDS. However, only 53 hospitalized patients were involved in their model development and validation. To obtain statistical significance a larger sample size is desired. Liang et al. used LASSO and logistic regression to develop a predictive risk scoring system called COVID-GRAM [43] based on 1590 patients. LASSO regression is used to extract 10 key risk factors out of 72 features, and the logistic regression was employed because of its interpretability. This model was able to predict a "severe" COVID-19 progression (defined by a composite of ICU admission, ventilation, and death), but did not have differentiation on which specific stage will the disease escalate to. Similar to Liang's work, Xie et al. performed multivariable logistic regression to predict mortality among patients [44]. These works demonstrate the capability of logistic regression in COVID-19 prognosis, which inspires us to integrate the logistic regression into our prediction pipeline.

Besides logistic regression, tree-based decision models have also been used to predict severe COVID-19 disease progression. Iwendi et al. applied random forest boosted by AdaBoost algorithm to predict the mortality of patients [45]. The model used the information collected by questionnaires but failed to leverage direct measurements like laboratory, vital, or radiological

findings. Similar to the AdaBoost approach, Yan et al. used XGBoost classifier to identify risk factors [46]. The importance of features in XGBoost is reflected by its cumulated use, meaning those more discriminative features tend to be used more often for splits (XGBoost continuously splits its internal nodes into sub-nodes until the sub-nodes are clean).

In summary, these statistical learning methods are more sophisticated than nomograms in terms of model complexity. However, in practice, these traditional statistical learning models demand much domain knowledge (e.g., for the SVM algorithm, the kernel function needs to be customized to fit non-linear patterns). Considering a novel disease like COVID-19, where existing knowledge is not as adequate, deep-learning models provide a purely data-driven alternative to fit complex patterns.

### 2.2.3 Deep-learning models

According to the Universal Approximation Theorem, a neural network can potentially approximate any continuous function [53]. This promise has drawn great enthusiasm in applying neural networks to different research areas, including COVID-19 prognosis. There are multiple deep-learning models developed for predicting COVID-19 progression. These models are especially useful in handling image-type input like CT and X-ray, where existing network architectures in computer vision can be adopted [49].

Meng et al. developed a model called De-COVID19-Net, a 3D densely connected convolutional neural network for predicting the survival of patients within a 14-day time window [47]. The CT images, together with sex, age, severity grade, and chronic disease were input into the 121-layer 3D-CNN network. Lee et al. adopted the recursive neural network (RNN) to predict severe outcomes of patients based on their historical electronic health records (EHR)

prior to hospital admission [48]. A novelty of this model is that instead of using the measurements taken after diagnosis, the proposed RNN model leverages the historical data prior to hospital admission. However, a resulting drawback is that a coming COVID-19 patient does not necessarily have their EHR records established, and thus does not have sufficient data for prediction.

Fakhfakh et al. proposed ProgNet, a combination of CNN and RNN to predict the severity of lung pathologies due to Covid-19 [49]. The network structure is nested, where each RNN unit contains a full CNN (i.e., for each time step, an RNN unit takes in the output of CNN). Although this model achieves decent accuracy, its feature space is confined to only radiology data: it did not use other clinical features which may complement the X-ray images. Furthermore, the RNN model requires an extensive number of x-ray images in time series, which is not accessible for most patients since they do not frequently take CT scans. In general, deep learning-based prognostic models tend to have high accuracy when training data are sufficient, but their performance can quickly degrade as the data become insufficient.

## 2.3 Research Gaps Addressed by This Thesis

As discussed above, a variety of methods have been developed to identify high-risk patients, providing an early alarm for severe disease progression. However, it is worth noting that these existing methods are coarse-grained in terms of model output: they predict either a single-stage outcome or the composition of multiple stages, without separated risk evaluations for each stage. There remains a critical need to develop a model predicting disease escalation in a fine-grained, to facilitate targeted therapy for patients at different risk levels.

To enable a fine-granularity prediction, this thesis integrates a multi-stage scheme into the prediction pipeline. The risks are evaluated upon a sequence of clinical events including ICU admission, mechanical ventilation, and mortality. Aligned with the multi-stage scheme, prediction is performed progressively, allowing model outputs to be updated whenever a new event is observed.

Compared with the existing deep-learning models, our model places more emphasis on interpretability: from the trained coefficients, clinicians can tell from the logistic regression model the contribution of each risk factor to final outcome. Such emphasis would be appreciable in clinical practice because an interpretable model is found more likely to be adopted by clinicians [35]. Besides, among the methods listed in Table 1, the use of deep learning did not bring significant improvement in accuracy but largely increased model complexity. For example, the De-COVID19-Net proposed by Meng et al. has up to 121 layers [47], and its 3-D convolutional structure further enhances the complexity. Such "black-box" models appear hard for clinicians to verify and rationalize, resulting in a slow model adopting rate. According to Nisha et al., if clinicians find a model understandable, they are more inclined to accept the outputs of a model [54]. In this sense, the traditional statistical learning approaches, such as multivariate logistical regression, are competitive in the clinical context due to their interpretability. Therefore, our multi-stage predictive model adopts logistic regression in its pipeline. The superiority of logistic regression is also verified by an experiment presented in Chapter 5, which compares the effectiveness of logistic regression with other prevalent methods.

Another uniqueness of this thesis is its focus on the greater Chicago area, which is not yet explored by the aforementioned studies. It is crucial to customize models for different geographic regions, supported by a study showing that the patients in different regions have

distinct characteristics [55]. For example, the patterns found from Chinese patients may not fit the patients in the United States. Moreover, even in the United States, different cities have a substantial difference in their population, considering the diverse race and ethnicity composition [56]. Therefore, although a number of models have been published for COVID-19 patients worldwide, it is still essential to developed models oriented at the Chicago population.

To address these research gaps, this thesis presents a multi-stage prognostic model for a diverse patient population in Chicago. This multi-stage model stratifies the patients into different risk levels, allowing healthcare providers to tailor therapy and prepare medical resources in advance. Practically, these patients are stratified in terms of clinical events, i.e., hospital admission, ICU admission, mechanical ventilation, and death. Then the laboratory measurements are presented in a multi-stage perspective and the triage models are subsequently built. Logistic regression is used to build the predictive model because this interpretable model can facilitate rapid clinical translation. Finally, the risk factors are extracted using two interpretable machine learning algorithms: 1) LASSO regression and 2) decision tree. When these two algorithms reach a consensus, the importance of risk factors is consolidated.

# CHAPTER 3

# PATIENT CHARACTERISTICS

Our analysis included 10123 COVID-19 patients for whom information on demographics and initial symptoms (e.g., cough, fever, muscle pain) was available. Of these patients, 1788 were admitted to the hospital; for these hospitalized patients, additional longitudinal information, such as lab test results and vital signs, was available. Among the hospitalized patients, 1076 had most of their lab measures and vitals collected on the day of admission. These patients' records were used to develop and validate a multi-stage prognostic framework and to identify top risk factors for severe disease progression.

**Data Sources**

Data was collected at Rush University Medical Center in Chicago, Illinois, and includes COVID-19 patients evaluated between 03/17/2020 and 08/07/2020. Patient and treatment information was obtained from queries against data warehouses populated from regular exports of clinical data stored in Rush's Epic electronic medical record (EMR) system. The study population included 10,123 patients with COVID-19, and their EMRs were retrospectively accessed to extract the patients' demographics (Table 2), laboratory findings, vitals (Table 3 and 4), and comorbidities (Table 5). The age of each patient was typically documented as a numerical value, but for 41 patients whose ages were documented by the text "90+", we assigned an age of 90 in our analysis.

### Variables

Variables investigated in our study include patient demographics, laboratory findings, vital signs, and comorbidities as documented in the EMR. Demographic variables include patients' age, sex, race, and smoking status. Laboratory findings include white blood count, absolute neutrophil count, absolute lymphocyte count, absolute monocyte count, neutrophils percentage, lymphocyte percentage, monocyte percentage, albumin, aspartate transaminase, alanine transaminase, d-dimer, red blood count, blood urea nitrogen, creatinine, hemoglobin, ferritin, C-reactive protein, lactate dehydrogenase, blood glucose, platelet count, and creatine phosphokinase. Vitals include oxygen saturation ($SpO_2$), body temperature, respiration rate, blood pressure, and pulse. Comorbidities included hypertension, type 2 diabetes, chronic kidney disease, pulmonary disease, and chronic ischemic heart disease.

## 3.1 Overall Demographic and Symptomatic Characteristics

The demographic information, symptoms, and smoking status of infected patients are presented in Table 2. With a median age of 40 ($75^{th}$ %ile 54), the patients were younger than those reported from China [1] and New York [6]. There were more infected females (53.55%) than males (46.45%). In terms of race, African Americans (32.92%) and Whites (27.27%) constituted the largest percentage of infected patients. Cough (70.14%), fever (46.86%), and shortness of breath (41.32%) were the most common symptoms. In terms of ethnicity, marginally more Hispanics or Latinos (50.88%) were infected than Non-Hispanics or Latinos (49.12%). People who had never smoked (76.42%) and former smokers (16.18%) accounted for most of the infected patients; Current smokers accounted for only 7.4%.

**Table 2 Demographics, symptoms, and smoking status of infected patients**

| Patient Characteristics | Overall |
|---|---|
| Age, median (IQR) | 40 (28–54) |
| Sex, N (%) (out of 10,120 patients whose sex was reported) | |
| Female | 5,419 (53.55%) |
| Male | 4,701 (46.45%) |
| First Race, N (%) (out of 8,208 patients whose first race was reported) | |
| African American | 2,702 (32.92%) |
| White | 2,238 (27.27%) |
| Asian | 140 (1.71%) |
| Native Hawaiian or Other Pacific Islander | 14 (0.17%) |
| American Indian or Alaska Native | 15 (0.18%) |
| Other | 3,099 (37.76%) |
| Ethnicity, N (%) (out of 9,379 patients whose first race was reported) | |
| Hispanic or Latino | 4,772 (50.88%) |
| Not Hispanic or Latino | 4,607 (49.12%) |
| Symptoms, N (%) (out of 5,499 patients who had symptom records) | |
| Cough | 3,857 (70.14%) |
| Fever | 2,577 (46.86%) |
| Shortness of Breath | 2,272 (41.32%) |
| Muscle Pain | 1,609 (29.26%) |

**Table 2 Continued**

| Patient Characteristics | Overall |
|---|---|
| Sore Throat | 1,239 (22.53%) |
| Loss of Smell | 1,079 (19.62%) |
| Smoking Status, N (%) (out of 6,074 patients whose smoking status was within the following categories) | |
| Never Smoker | 4642 (76.42%) |
| Former Smoker | 983 (16.18%) |
| Current Every-day Smoker | 295 (4.86%) |
| Current Some-days Smoker | 154 (2.54%) |

The ages of the infected patients are presented in Figure 2(a), which represents data from 03/17/2020 and after. The interquartile range (IQR) of age is represented by the shaded region. The median age decreased from 46 (in week 0) to 37 (in week 20), indicating the spread of disease to a younger population. The number of clinical events, including hospital admissions, ICU admissions, and mortalities, are presented in Figure 2(b). The peak of hospital admissions was reached at week 5; however, the peak of ICU admissions was reached at week 1. The peak of mortality was reached at week 7. Week 7 (which ended on 05/05/2020) was a turning point in the number of hospital admissions, indicating a reduction in transmission of COVID-19 in the Chicago area. Then, the second wave of infections is suggested by the increase in hospital admissions from week 14 to week 19.


(a) (b)

**Figure 2 Time-series plot on weekly granularity (data from 03/17/2020 onwards): (a) Age of infected population (the interquartile range (IQR) of age is marked with the shaded region), (b) Number of hospital admissions, ICU admissions, and deaths.**

## 3.2 Laboratory Measures and Vital Signs of Hospitalized Patients

Table 3 presents the laboratory measures and vitals of hospitalized patients on the day of admission. Among all the investigated laboratory measures, the lymphocyte percentage (median 16.3, IQR 10.6–23.2) and albumin (median 3.3, IQR 3.0–3.7) were lower than the normal range, while ferritin (median 766.2, IQR 325.9–1643.0), d-dimer (median 0.9, IQR 0.5–2.5), C-reactive protein (median 112.1, IQR 54.5–192.8), lactate dehydrogenase (median 404.0, IQR 304.0–544.0), and blood glucose (median 122.0, IQR 103.0–175.0) were higher than the normal range. Among vital signs, the respiratory rate (median 20.6, IQR 18.6–24.0)) was slightly higher than the normal range.

**Table 3 Laboratory measures and vital signs of hospitalized patients on their dates of admission**

| Measures | Median (IQR) |
|---|---|
| Laboratory tests | |
| White Blood Count (K/Ul, Low: 4.00, High: 10.00) | 7.3 (5.4–10.0) |
| Neutrophil Absolute Count (K/Ul, Low: 1.84, High: 7.80) | 5.2 (3.6–7.7) |
| Lymphocyte Number (K/Ul, Low: 0.72, High: 5.20) | 1.1 (0.8–1.5) |
| Monocyte Number (K/Ul, Low: 0.12, High: 1.00) | 0.5 (0.3–0.7) |
| Neutrophils Percent (%, Low: 46.0, High: 78.0) | 74.6 (65.9–81.7) |
| Lymphocyte Percent (%, Low: 18.0, High: 52.0) | 16.3 (10.6–23.2) |
| Monocyte % (%, Low: 3.0, High: 10.0) | 7.0 (4.9–9.2) |
| Albumin (G/Dl, Low: 3.5, High: 5.0) | 3.3 (3.0–3.7) |
| Sgot (U/L, Low: 3, High: 44) | 38.0 (25.0–59.0) |
| Sgpt (U/L, Low: 0, High: 40) | 30.0 (18.0–49.0) |
| Red Blood Count (M/Ul, Low: 4.00, High: 5.20) | 4.4 (3.9–4.9) |
| Urea Nitrogen (Mg/Dl, Low: 8, High: 21) | 14.0 (10.0–24.0) |
| Creatinine (Mg/Dl, Low: 0.65, High: 1.00) | 1.0 (0.8–1.4) |
| Hemoglobin (G/Dl, Low: 12.0, High: 16.0) | 12.9 (11.1–14.2) |
| Ferritin (Ng/Ml, Low: 12, High: 260) | 766.2 (325.9–1643.0) |
| D-Dimer (Mg/L Feu, Low: 0.00, High: 0.60) | 0.9 (0.5–2.5) |
| C-Reactive Protein (Mg/L, Low: 0.0, High: 8.0) | 112.1 (54.5–192.8) |

**Table 3 Continued**

| Measures | Median (IQR) |
|---|---|
| Lactate Dehydrogenase (U/L, Low: 110, High: 240) | 404.0 (304.0–544.0) |
| Glucose, Blood (Mg/Dl, Low: 60, High: 99) | 122.0 (103.0–175.0) |
| Platelet Count (K/Ul, Low: 150, High: 399) | 218.0 (170.0–281.0) |
| Creatine Phosphokinase (U/L, Low: 10, High: 205) | 128.0 (66.0–332.5) |
| Vitals | |
| $SpO_2$ (%, normal range 95–100) | 95.8 (94.2–97.4) |
| Temperature (°F, normal range 97–99) | 98.8 (98.0–99.8) |
| Respiration Rate (Breaths per minute, normal range 12–20) | 20.6 (18.6-24.0) |
| Pulse (Beats per minute, normal range 60–100) | 90.1 (80.0–101.7) |

# CHAPTER 4
# A MULTI-STAGE SCHEME TO MODEL DISEASE PROGRESSION

## 4.1 An Event-Based Multi-Stage Stratification Scheme

In order to discretize the disease progression continuum, we defined the stages of disease progression according to the occurrence of clinical events requiring increasing levels of medical resources. We defined (1) *hospitalization*, (2) *admission to ICU*, and (3) *mechanical ventilation* as advancing stages of disease progression, and (4) *death* as the terminal stage of COVID-19 disease. Figure 3 presents the transition of COVID-19-infected individuals across those stages.

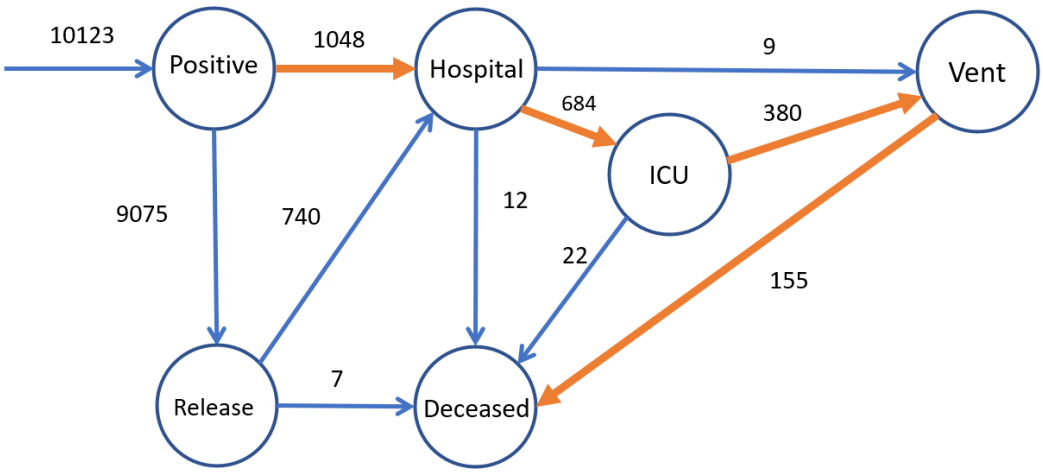



**Figure 3 Transitions of patients among clinical stages.**

As shown in Figure 3, a total of 10,123 individuals were tested positive for COVID-19. Among them, 1,788 required hospitalizations. The patients who suffered severe disease progression were treated in the ICU; some also received ventilatory support, depending on the severity of their disease. Over 38% (684) of the patients admitted to the hospital were treated in

the ICU, and of those in the ICU, over 55% (380) received ventilatory support. Finally, over 40% (155) of the patients who received ventilation died.

## 4.2 Laboratory, Comorbidity, and Radiographic Findings Under the Multi-Stage Scheme

With the staging scheme defined, Table 4 presents the laboratory measures at different stages. From the hospitalization stage to ICU stage to the ventilation stage, all laboratory measures have their median values change monotonically except for Monocyte percentage and Lymphocyte percentage.

Figure 4 presents the temporal changes in laboratory features, covering 15 days after hospital admission. To reduce the appearance of the same patients in multiple groups, we classified the patients into four groups: hospitalized but not in the ICU; in the ICU but not ventilated; ventilated but not deceased; and deceased. The albumin level of all patient groups decreased from day 0 to day 7, indicating a general catabolic state, not uncommon in hospitalized patients in general. During that time, the albumin levels of the ventilated patients who ultimately survived behaved much like those patients who subsequently died. However, the albumin levels of the eventual survivors then began to increase, while the albumin levels of those who subsequently died continued to decrease until day 11. The blood urea nitrogen level of the patients who eventually died kept increasing until death and was generally higher than the level found in other groups of patients. The lymphocyte percentages of ICU-but-never-ventilated patients tended to move from the abnormal range (<18%) to the normal range, while the lymphocyte percentages of patients who eventually died tended to worsen. This correlation means that the lymphocyte percentage can be used as an indicator of a patient's condition. A rising lymphocyte percentage indicates a recovery trend in a COVID-19 patient, while a

decreasing lymphocyte percentage indicates a worsening condition. As for the red blood count, the initial values for patients who later died were in roughly the same range as those of survivor groups, but the value for patients who died showed a sharper decrease later compared with the survivors. Thus, the initial value for the red blood count may not be a risk factor for mortality prediction, but its downward trend can be used to forecast deterioration.

We investigated five comorbidities for COVID-19 patients. Ranked by their rates of appearance among the hospitalized patients, they include hypertension (60.29% among the hospitalization patients), overweight or obese condition (57.1%), type 2 diabetes (43.12%), chronic kidney disease (23.04%), and chronic ischemic heart disease (18.06%), as shown in Table 5. As COVID-19 progressed from the hospitalization stage to the ICU stage to eventual death, the percentage of patients who had chronic kidney disease or chronic ischemic heart disease constantly increased, indicating that these two comorbidities are significant risk factors for severe disease progression. In addition, the rates of hypertension and type 2 diabetes were also higher among deceased patients than among the total hospitalized population. The patients who had none of these five comorbidities and presented as relatively healthy accounted for very few (14.65%) of the hospitalized patients and even fewer of the deceased patients (10.71%), meaning that an originally healthy individual is far less likely to suffer serious effects due to COVID-19. These findings indicate that the studied comorbidities are risk factors for mortality.

**Table 4 Initial laboratory measures of the patients at different stages**

| Measures | Hospitalization Period | ICU Period | Ventilation Period |
|---|---|---|---|
| White Blood Count (K/Ul, Low: 4.00, High: 10.00) | 7.6 (5.8-10.4) | 10.1 (7.2-13.3) | 12.6 (9.8-15.2) |
| Neutrophils Percent (%, Low: 46.0, High: 78.0) | 70.9 (63.7-77.7) | 76.2 (69.4-82.0) | 80.2 (74.8-84.2) |
| Lymphocyte Percent (%, Low: 18.0, High: 52.0) | 18.0 (12.2-24.4) | 12.9 (8.9-18.4) | 9.7 (7.1-13.2) |
| Monocyte % (%, Low: 3.0, High: 10.0) | 7.7 (5.8-9.7) | 6.6 (5.0-8.4) | 5.7 (4.1-7.4) |
| Albumin (G/Dl, Low: 3.5, High: 5.0) | 2.9 (2.5-3.4) | 2.4 (2.0-2.9) | 2.0 (1.7-2.4) |
| Sgot (U/L, Low: 3, High: 44) | 39.0 (25.0-60.4) | 49.7 (32.0-79.4) | 62.4 (39.7-98.5) |
| Sgpt (U/L, Low: 0, High: 40) | 33.8 (19.2-58.7) | 40.6 (22.5-69.0) | 48.0 (28.9-81.0) |
| Red Blood Count (M/Ul, Low: 4.00, High: 5.20) | 4.2 (3.7-4.6) | 3.9 (3.3-4.4) | 3.6 (3.1-4.0) |
| Urea Nitrogen (Mg/Dl, Low: 8, High: 21) | 15.4 (10.6-26.0) | 23.7 (14.4-41.4) | 32.7 (21.6-48.3) |
| Creatinine (Mg/Dl, Low: 0.65, High: 1.00) | 0.9 (0.7-1.3) | 1.1 (0.8-2.2) | 1.4 (0.8-2.8) |

**Table 4 Continued**

| Measures | Hospitalization Period | ICU Period | Ventilation Period |
|---|---|---|---|
| Hemoglobin (G/Dl, Low: 12.0, High: 16.0) | 12.1 (10.4-13.4) | 11.2 (9.3-12.7) | 10.4 (8.8-11.7) |
| Ferritin (Ng/Ml, Low: 12, High: 260) | 776.7 (368.1-1702.3) | 1212.8 (556.2-2334.4) | 1555.1 (895.8-2466.9) |
| D-Dimer (Mg/L Feu, Low: 0.00, High: 0.60) | 1.1 (0.6-3.4) | 3.0 (1.0-6.4) | 4.6 (2.3-10.4) |
| C-Reactive Protein (Mg/L, Low: 0.0, High: 8.0) | 102.3 (52.5-164.8) | 153.5 (94.9-224.3) | 200.8 (129.6-261.3) |
| Lactate Dehydrogenase (U/L, Low: 110, High: 240) | 390.0 (291.2-525.2) | 506.9 (384.0-657.5) | 576.1 (442.9-722.2) |
| Glucose, Blood (Mg/Dl, Low: 60, High: 99) | 120.8 (101.0-166.9) | 143.8 (112.3-194.3) | 169.7 (132.0-215.0) |
| Platelet Count (K/Ul, Low: 150, High: 399) | 242.5 (186.4-311.4) | 246.7 (184.1-319.6) | 247.6 (182.0-322.4) |
| Creatine Phosphokinase (U/L, Low: 10, High: 205) | 151.0 (66.0-360.0) | 223.1 (100.2-594.8) | 342.5 (160.2-991.5) |

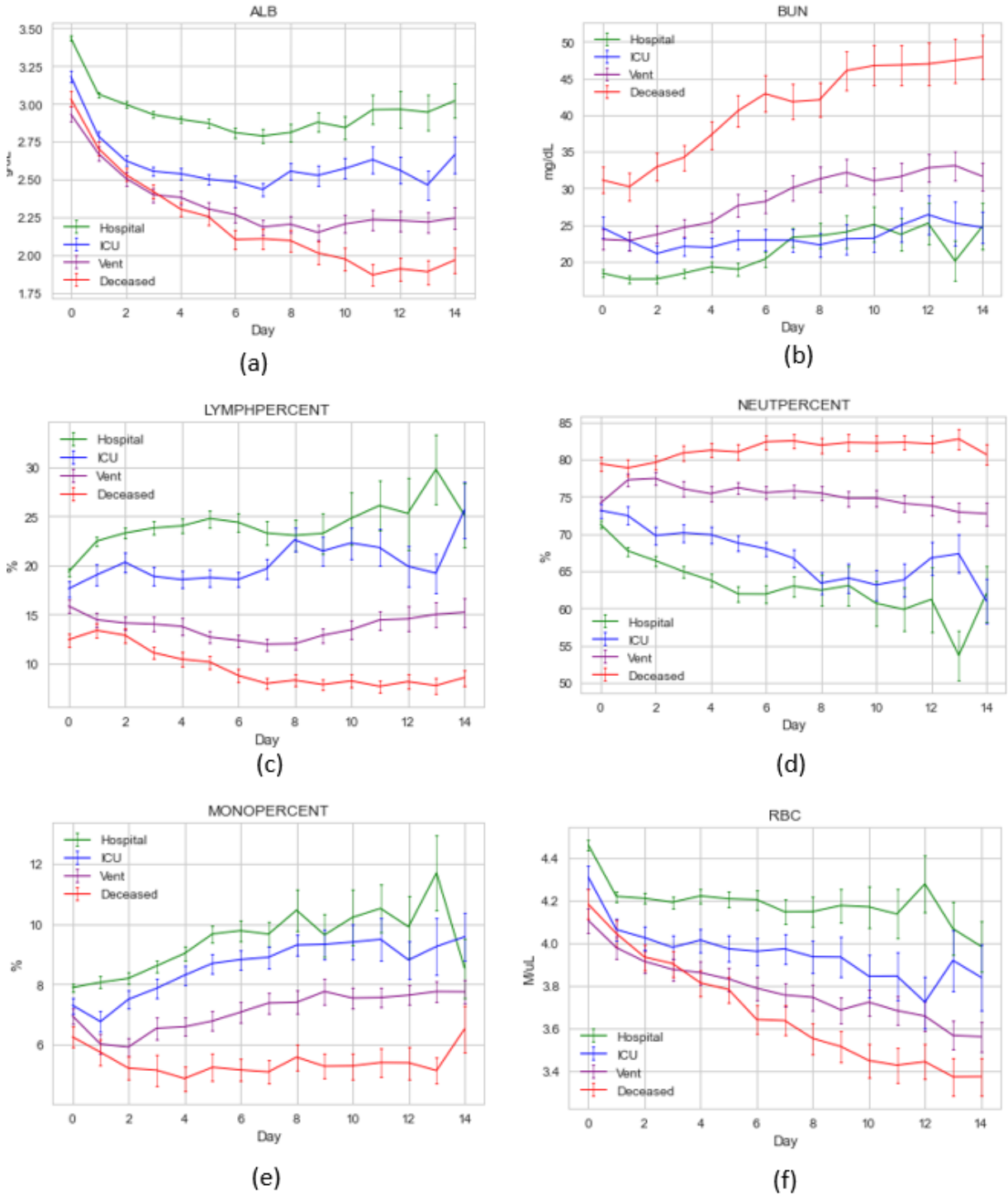



**Figure 4 Temporal changes of laboratory features from the day of hospital admission to 15 days in hospital: (a) albumin level, (b) blood urea nitrogen level, (c) lymphocyte percentage, (d) neutrophil percentage, (e) monocyte percentage, (f) red blood count.**

**Table 5 Comorbidities among the COVID-19 patients**

| Comorbidity | COVID-19-Positive | Hospitalized | p-value (Hospitalized - Entering ICU) | Entering ICU | p-value (Entering ICU-Ventilated) | Ventilated | p-value (Ventilated - Deceased) | Dead |
|---|---|---|---|---|---|---|---|---|
| Hypertension | 2,225 (21.99%) | 1,078 (60.29%) | <0.001 | 468 (68.42%) | 0.1621 | 282 (72.49%) | 0.0763 | 139 (70.92%) |
| Overweight or obese | 2,033 (20.09%) | 1,021 (57.1%) | 0.0135 | 428 (62.57%) | 0.024 | 270 (69.41%) | <0.001 | 105 (53.57%) |
| Type 2 diabetes | 1,422 (14.05%) | 771 (43.12%) | <0.001 | 351 (51.32%) | 0.5498 | 207 (53.21%) | 0.1124 | 100 (51.02%) |
| Chronic kidney disease | 515 (5.09%) | 412 (23.04%) | 0.003 | 197 (28.8%) | 0.9974 | 112 (28.79%) | 0.7198 | 67 (34.18%) |
| Chronic ischemic heart disease | 437 (4.32%) | 323 (18.06%) | 0.0152 | 153 (22.37%) | 0.7687 | 84 (21.59%) | 0.826 | 46 (23.47%) |
| Without any of the above 5 comorbidities | 6,693 (66.14%) | 262 (14.65%) | <0.001 | 62 (9.06%) | 0.289 | 28 (7.2%) | 0.325 | 21 (10.71%) |

A total of 989 patients received X-Ray or CT scanning on hospital admission day. Key findings from radiological reports were presented in Figure 5. “Lung Opacity” was the most prevalent findings, with over 70% occurrence rate among hospitalized patients. “Lung Opacity” was most widely observed among those who ultimately deceased. The appearing rate of “Enlarged Cardiomediastinum” and “Atelectasis” steadily increased as the outcome worsens, indicating their positive correlations with severe disease progressions. Finally, “Edema” (3.1%) and “Pneumothorax” (0.4%) have comparatively low occurrence rates at hospital admission.

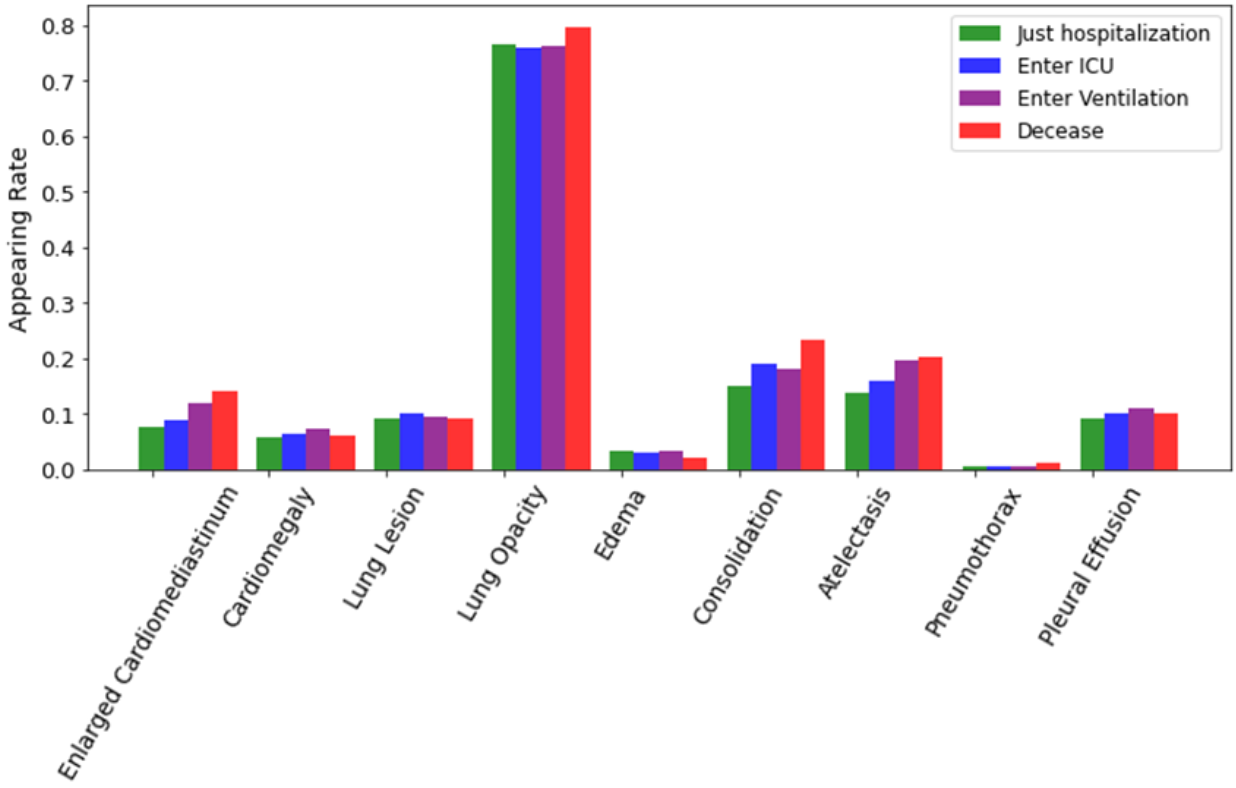



**Figure 5 Appearance rate of radiological findings among the patients with different outcomes. (The radiological findings were extracted from the radiological reports obtained at hospital admission day, using the labeling tool described in** [57]**.)**

# CHAPTER 5

# MODEL DEVELOPMENT AND EVALUATION

## 5.1 Classifiers for Initial and Progressive Triages

### 5.1.1 Variables Used in the Triaging Scheme

We used a subset of the aforementioned variables to develop our triaging schemes. Those variables were selected using a two-step procedure: 1) First, we selected a number of variables used in previous COVID-19 triaging studies from China and other states in the United States, such as neutrophil-to-lymphocyte ratio [58], [59], albumin level [60], and creatinine level [61]. 2) Second, we selected a subset among those variables that was highly represented (measured in over 65% of the patients at their hospital admission day) in the Rush University Medical Center dataset. The selected variables include age, race, sex, neutrophil-to-lymphocyte ratio; neutrophil, lymphocyte, and monocyte percentages; white blood, red blood, and platelet counts; and the levels of blood glucose, blood urea nitrogen, creatinine, albumin, aspartate transaminase, alanine transaminase, hemoglobin, and $SpO_2$.

### 5.1.2 Triaging Schemes

In this study, we developed two different triaging schemes using the variables and disease stages described previously. We utilized variables measured at both hospital admission (referred to as baseline triage) and current disease stage (referred to as progressive triage) for the two scenarios, as shown in Figure 6.

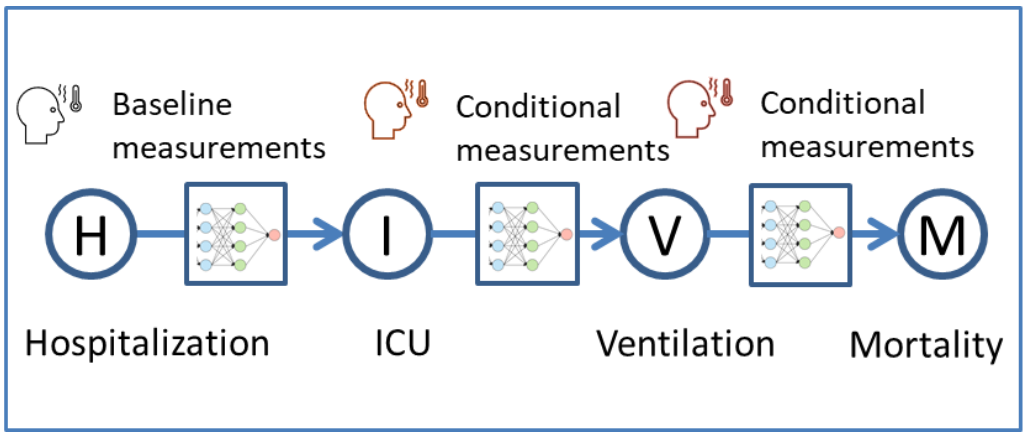



**Figure 6 Illustration of using baseline measurements and conditional measurements for multi-stage prognosis.**

**Baseline triage:** Baseline triaging scheme consists of a set of binary classifiers predicting the likelihood of reaching each disease progression stage from baseline, i.e., whether a patient will be admitted to the ICU, be ventilated, or die, respectively.

**Progressive triage**: The progressive scheme considers the current stage of the disease and predicts escalation to advanced stages. This scheme performs triaging in three steps: a) when a patient is admitted to the hospital, a binary classifier predicts whether the patient will recover or progress to a more critical stage (i.e., ICU, mechanical ventilation, or death); b) when a patient is admitted to the ICU, a binary classifier predicts whether the patient will recover or progress to a more critical stage (i.e., mechanical ventilation or death); and c) when a patient is treated using a mechanical ventilator, a binary classifier predicts whether the patient will recover or progress to death. In other words, given the current event (hospital admission, ICU admission, or mechanical ventilation), the system predicts whether the disease will diminish or progress.

## 5.2 Classifiers Training and Evaluation

### 5.2.1 Classifier Training and Evaluation

We chose logistic regression as the classifier in both scenarios because an easily interpretable model can help facilitate rapid clinical translation [62]. The logistic regression classifiers were trained and validated using a standard 10-fold cross-validation approach [63]. The accuracy of prediction was evaluated using the area under the receiver-operator characteristic curve (AUC). We selected a cutpoint by fixing the sensitivity higher than 0.7 (to guarantee the detection of positive patients) and meanwhile maximizing the specificity. The processes of cross-validation and cutpoint selection are visualized in Figure 7.

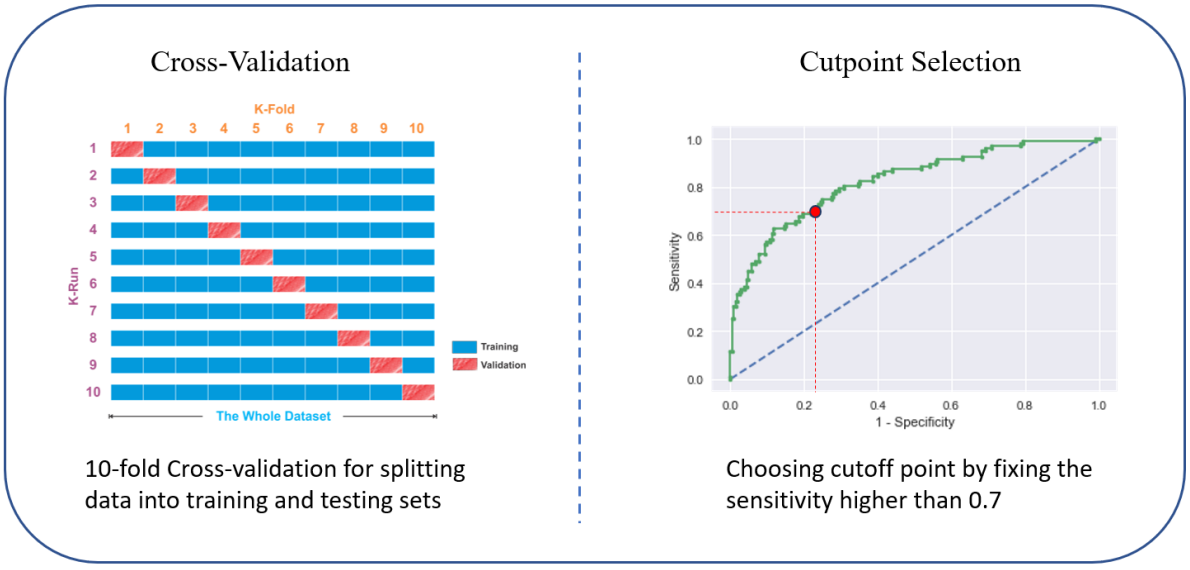



**Figure 7 Training and evaluation of the classifier.**

### 5.2.2 Preliminary Experiment

To verify the effectiveness of logistic regression for predicting severe COVID-19 outcomes, a preliminary experiment is conducted to compare logistic regression to other

prevalent machine learning methods, including decision tree, random forest, SVM, and neural network. The task is to predict the ventilation demands among hospitalized patients, and the model performance is evaluated using 10-fold cross-validation specified above. The experiment results are summarized in Table 6. In comparison, logistic regression yields the highest AUC, accuracy, and F-1 Score, which demonstrates the superiority of using logistic regression to predict severe COVID-19 outcomes.

**Table 6 Comparison of machine learning models for predicting ventilation demand**

| Model | AUC | Accuracy | F-1 Score |
|---|---|---|---|
| Logistic Regression | 0.819 | 0.784 | 0.620 |
| Decision Tree | 0.744 | 0.771 | 0.592 |
| Random Forest | 0.780 | 0.777 | 0.617 |
| Support Vector Machine | 0.751 | 0.684 | 0.02 |
| Neural Network | 0.728 | 0.757 | 0.503 |

### 5.2.3 Baseline triage

Table 7 summarizes the results of baseline triage. In comparison, a more severe outcome was more predictable (resulting in a higher AUC) at the baseline. The prediction for mortality has the highest AUC of 0.803 (95% CI: 0.752 - 0.853). With a cutpoint at 0.095, the model was able to identify 71.0% of eventually deceased patients at their hospital admission, meanwhile correctly predicting 73.9% of those finally survived.

**Table 7 Results of Initial triage for ICU admission, ventilation admission, and mortality**

| Output classes | Whether admitted into ICU | Whether ventilated | Whether eventually die |
|---|---|---|---|
| AUC (95% CI) | 0.736 (0.718 - 0.754) | 0.767 (0.727 - 0.807) | 0.803 (0.752 - 0.853) |
| Optimal Cutpoint | 0.365 | 0.19 | 0.095 |
| Confusion Matrix | [392 231<br>133 320] | [542 290<br>70 174] | [716 253<br>31 76] |
| Sensitivity | 0.706 | 0.713 | 0.710 |
| specificity | 0.629 | 0.651 | 0.739 |

### 5.2.4 Progressive triage

Table 8 shows the results of progressive prediction using conditional features (i.e., classification was performed using the features gathered during the current event), which resulted in AUC values of 0.738 (95 %CI: 0.703 – 0.773) for the ICU admission prediction, 0.710 (95 %CI: 0.667 – 0.753) for the ventilation admission prediction, and 0.642 (95 %CI: 0.550 - 0.733) for mortality prediction. In comparison, the baseline triage above achieves higher accuracy, especially for mortality prediction, where the baseline triage has an AUC value 25.1% higher than that obtained using conditional features.

**Table 8 Results of progressive prediction using conditional features**

| Output classes | (Recover at the hospital) vs (Be admitted into ICU or beyond) | (Recover at ICU) vs (Be ventilated or beyond) | (Recover at ventilators) vs (Decease) |
|---|---|---|---|
| AUC (95% CI) | 0.738 (0.703 – 0.773) | 0.710 (0.667 – 0.753) | 0.642 (0.550 - 0.733) |
| Optimal Cutpoint | 0.37 | 0.59 | 0.335 |
| Confusion Matrix | [382 238<br>134 322] | [111 69<br>79 186] | [52 61<br>26 62] |
| Sensitivity | 0.706 | 0.702 | 0.705 |
| specificity | 0.616 | 0.617 | 0.460 |

## 5.3 Converting Model Outputs to Predict the Most Severe Stage

Next, we sought to identify the most severe clinical stage a patient is likely to reach using baseline features. Here we discuss a simple strategy to convert the results of the baseline prognostic model for identifying the most severe stage.

By outputting the individual probability of entering each stage, the baseline triage model provides an opportunity to anticipate the ultimate stage of escalation. We discuss here two simple strategies to convert the results of baseline triage for predicting the severest stage.

The first strategy is to take the stage with the largest probability (normalized by the prior probability of that stage) to be the severest one, expressed as the following equation:

$$severest = \underset{S\in\{Hosp,ICU,Vent,Death\}}{\operatorname{argmax}} \left(\frac{1-p_{ICU}}{\pi_{Hosp}}, \frac{p_{ICU}}{\pi_{ICU}}, \frac{p_{Vent}}{\pi_{Vent}}, \frac{p_{Death}}{\pi_{Death}}\right)$$

where $p_{ICU}$, $p_{Vent}$, and $p_{Death}$ are the output probabilities for ICU, ventilation, and mortality, respectively. Because the output probabilities of logistic regression are not equally distributed among these three classifiers, to make these output probabilities comparable, we equalized them by removing the tails and stretching the remainders to cover from 0 to 1. The result is presented in Table 9, where the model was able to detect those ultimately ventilated patients with a sensitivity of 0.619 and those ultimately deceased patients with a sensitivity of 0.636.

**Table 9 Conversion to the severest stage using the argmax of probabilities (normalized by prior probability of that stage)**

| Severest stage | Hospitalization | ICU | Ventilation | Mortality |
|---|---|---|---|---|
| Confusion Matrix | [419 37<br>464 156] | [840 34<br>186 16] | [559 370<br>56 91] | [665 304<br>39 68] |
| Sensitivity | 0.252 | 0.079 | 0.619 | 0.636 |
| specificity | 0.919 | 0.961 | 0.602 | 0.686 |

The second strategy is to progressively convert the output probabilities to the stages, by comparing the output probability to the cutpoint and deciding the escalation to the next stages. Such a decision process is visualized in Figure 8. For example, given a patient's ICU probability is higher than the cutpoint, he will be considered for ventilation and the output probability of the ventilation classifier will be compared with the cutpoint for ventilation. If the probability is not higher than the cutpoint, he will be considered not escalating to the ventilation stage, thus his severest stage remains at ICU admission. The flexibility of this strategy lies in the tunable cutpoints. When using the optimal cutpoints specified in the Table 7, the model is able to detect those merely hospitalized patients with a sensitivity of 0.629 and eventually deceased patients

with a sensitivity of 0.635, but the intermediate stages are not as differentiable, as shown in Table 10.

**Table 10 Conversion to the severest stage using the forward translation logic without tuning cutpoints**

| Severest stage | Hospitalization | ICU | Ventilation | Mortality |
|---|---|---|---|---|
| Confusion Matrix | [322 134<br>230 390] | [783 91<br>174 28] | [803 126<br>109 38] | [768 201<br>39 68] |
| Sensitivity | 0.629 | 0.138 | 0.258 | 0.635 |
| specificity | 0.706 | 0.896 | 0.864 | 0.792 |

Then we tuned the cutpoints to allow more ICU and ventilated patients to be correctly detected, which improves the sensitivity by 0.248 for ICU prediction and by 0.102 for ventilation prediction, as shown in Table 11.

**Table 11 Conversion to the severest stage using the forward translation logic without tuning cutpoints**

| Severest stage | Hospitalization | ICU | Ventilation | Mortality |
|---|---|---|---|---|
| Confusion Matrix | [322 134<br>230 390] | [783 91<br>174 28] | [803 126<br>109 38] | [768 201<br>39 68] |
| Sensitivity | 0.629 | 0.138 | 0.258 | 0.635 |
| specificity | 0.706 | 0.896 | 0.864 | 0.792 |

It is observed that tuning cutpoints can favor the prediction for those target stages, but at the cost of the prediction performance for other stages. In reality, such conversion will be done at clinicians' discretion. The involvement of clinicians not only helps in the selection of cutpoints, but also provides a more sophisticated decision process using their expert knowledge. The initial triage models proposed in this thesis provide references for their decision.

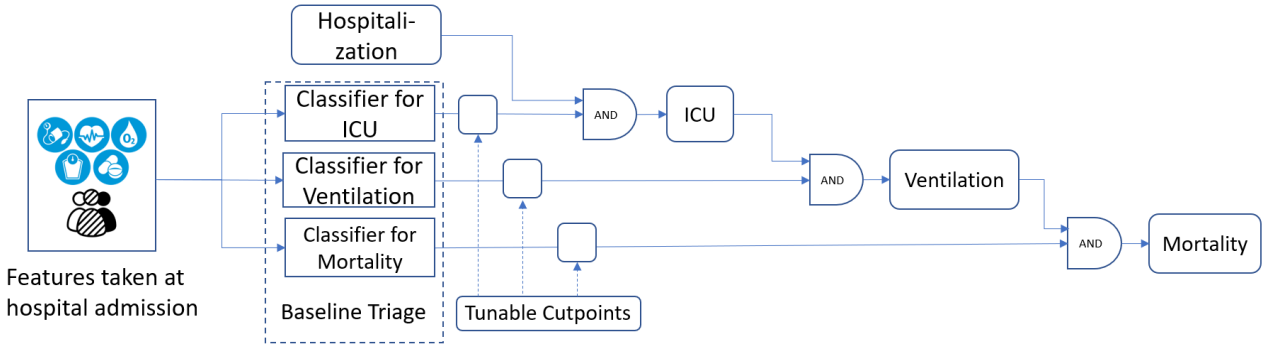



**Figure 8 Progressively convert the output probabilities to the stages by comparing the output probabilities to thresholds and deciding the escalation to the next stages.**

# CHAPTER 6
# RISK FACTOR ANALYSIS

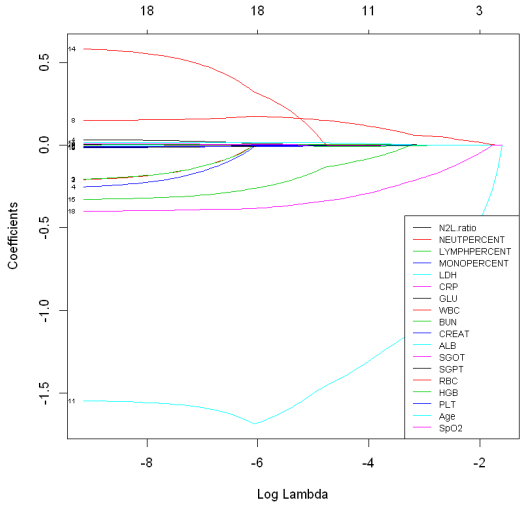
In this chapter, two prevalent machine learning models, LASSO regression and decision tree, are used to identify the key risk factors for severe disease progression.

## 6.1 Risk Factor Analysis Using LASSO Regression

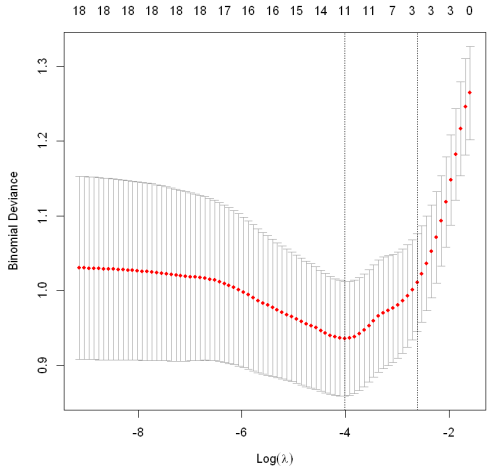
LASSO regression [64] was used to select the most discriminative variables for predicting the demand for mechanical ventilation. The "glmnet" packet in R was used to compute the results [65]. A total of 18 variables were entered in an L1-norm LASSO regression. They include age, neutrophil-to-lymphocyte ratio; neutrophil, lymphocyte, and monocyte percentages; white blood, red blood, and platelet counts; and the levels of lactate dehydrogenase, C-reactive protein, blood glucose, blood urea nitrogen, creatinine, albumin, aspartate transaminase, alanine transaminase, hemoglobin, and $SpO_2$. As the regularization term (typically denoted as $\lambda$) grows large, only the most important features are left with nonzero coefficients.

Figure 9 (a) shows the trace of coefficients as the $\lambda$ grows large. The coefficients of investigated features turn to 0 sequentially. Among all the features, albumin is the last one that turns to 0, meaning that albumin is the most discriminative feature selected by LASSO, followed by $SpO_2$ and white blood count. The significance of albumin is also indicated by the magnitudes of coefficients; the coefficient of albumin is consistently larger than that of any other feature. Albumin can be considered a general measure of an individual's overall health. Figure 9(b) shows the trace of binomial deviance, a type of misclassification error [66]. As the regularization

parameter λ becomes large, the binomial deviance first decreases and then increases. The axis above the figure shows the number of nonzero coefficients at a particular λ value. The minimal deviance is achieved with 12 features. They include age, platelet count, white blood count, neutrophil-to-lymphocyte ratio, lymphocyte percentage, and the levels of lactate dehydrogenase, C-reactive protein, blood glucose, blood urea nitrogen, albumin, hemoglobin, and $SpO_2$.




(a)




(b)

**Figure 9 LASSO regression for feature selection. (a) Trace of coefficients of the 18 baseline features. (b) Binomial deviance using 10-fold cross-validation, indicating the variation in misclassification error with different levels of regularization.**

## 6.2 Risk Factor Analysis Using Decision Tree Algorithm

In addition to performing a LASSO regression, we also constructed a decision tree [67] for the investigated variables, which provides another perspective that can be used to identify key risk factors. Using the Gini index as the impurity metric [68], the decision tree iteratively splits the current data into two branches. By definition, the variable used for splitting the root node is the most discriminative factor. Furthermore, the key risk factors tend to gather at the high-layer nodes near the root. The key risk factors selected by the LASSO regression and decision tree algorithms were further compared and entered in the logistic regression model for predicting ventilator demands.

To identify the most discriminative features, we used the decision tree algorithm with the Gini index. Figure 10 shows the result of the decision tree with max layer = 3. The first split is made on albumin = 2.65 g/dl. For the patients with albumin lower than 2.65 g/dl, the second split is made on the white blood count = 6.835 k/ul, indicating that the patients with low albumin and high white blood count are more likely to require mechanical ventilation. In contrast, for the patients with albumin higher than 2.65 g/dl, the second split is made on $SpO_2$ = 92.739%, indicating that the patients with high albumin and high $SpO_2$ values are less likely to need ventilatory support. The top three features selected by the decision tree algorithm exactly match those selected by LASSO regression, highlighting the significance of these features. In clinical practice, this compact set of features may be used to efficiently triage COVID-19 patients.

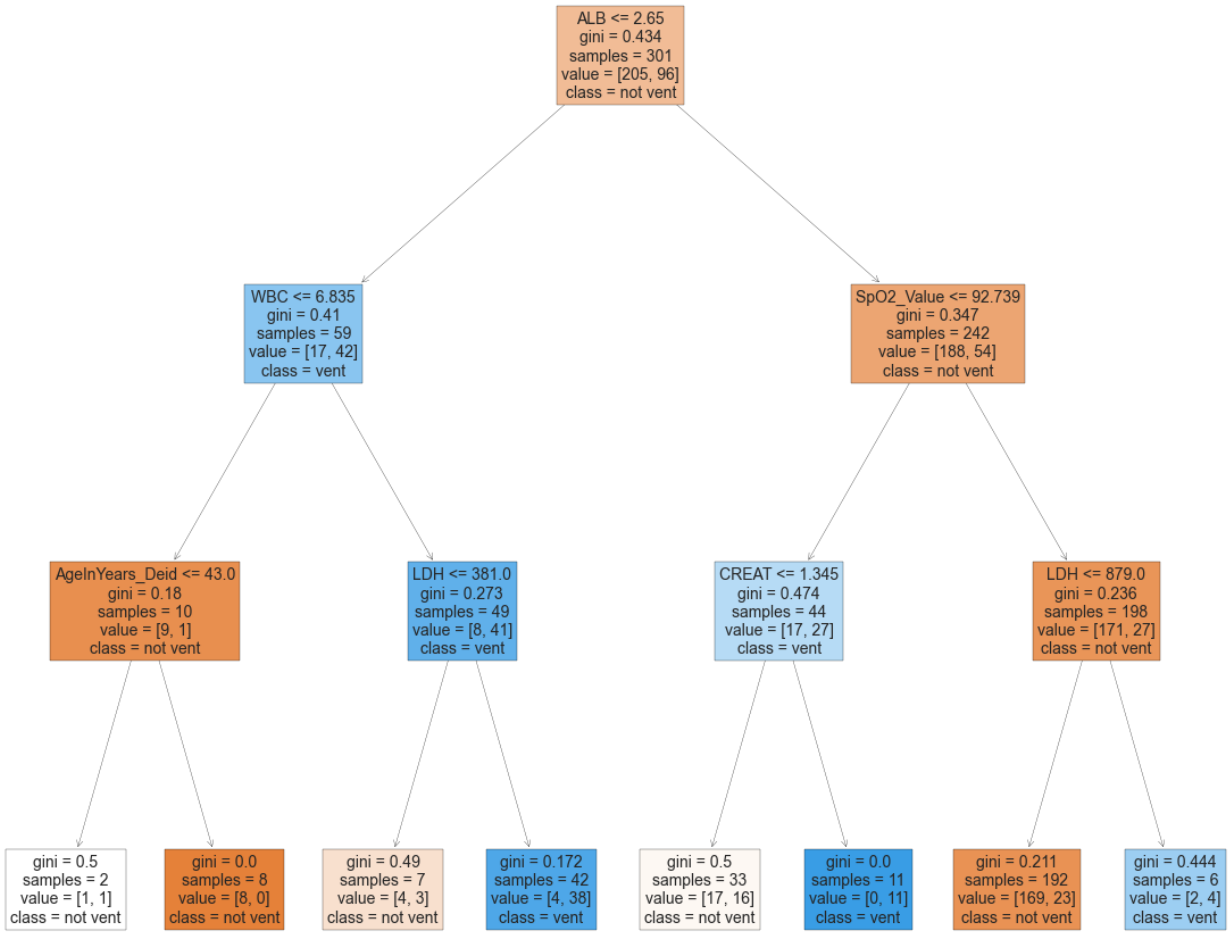



**Figure 10 Decision tree for distinguishing between patients who did and did not require ventilation.**

# CHAPTER 7

# CONCLUSION AND FUTURE WORK

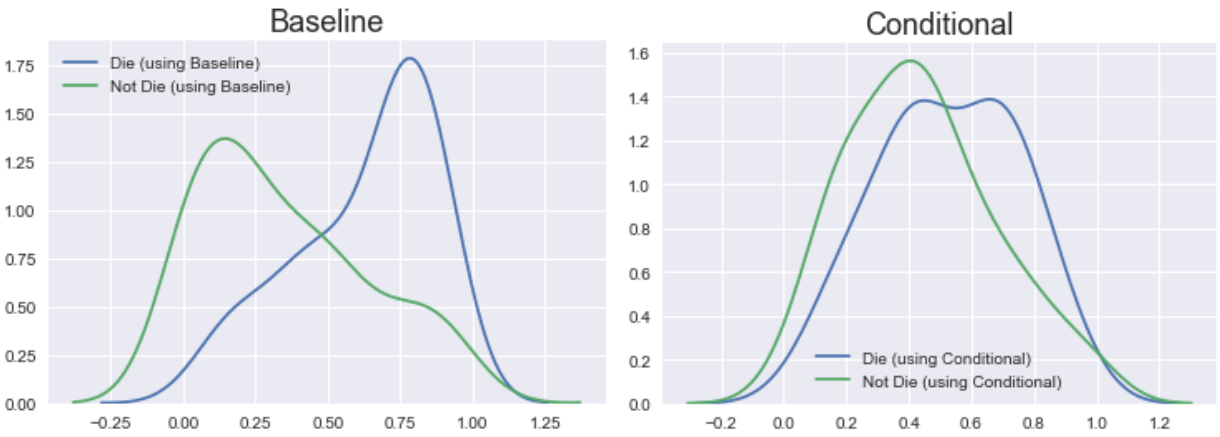
This thesis presents a novel multi-stage prognostic scheme for predicting COVID-19 disease progression. We characterized the stratification of COVID-19 patients in terms of clinical events and presented laboratory measures specific to each stratum, thereby making it easier to understand and track disease progression. Based on the stratification, we developed a multi-stage prognostic framework for predicting the probabilities of different outcomes of COVID-19 patients at both initial and progressive triage. We then used LASSO regression and decision tree models to identify several risk factors for the deterioration of patient health. This research and the resulting model establish the feasibility of an early triage tool that can predict the clinical course of COVID-19 at a subject-specific level, thus allowing precise allocation of medical resources for those in need.

In comparison, using baseline triage achieves higher accuracy than using conditional features, especially for mortality prediction. To ensure that it is the feature collection time that makes a difference (instead of distinct participants in those two cohorts), we select 159 patients who have lab tests on both hospital admission day and intubation day to build classifiers. The prediction results of these classifiers are presented in Table 12. The AUC using baseline features is 23.6% higher than that using conditional features. Besides the AUC metric, the prediction with baseline features has all PPV, NPV, sensitivity, and specificity metrics higher than the prediction with conditional features.

**Table 12 A comparison between using baseline features and conditional features among the patients who have both types of features documented**

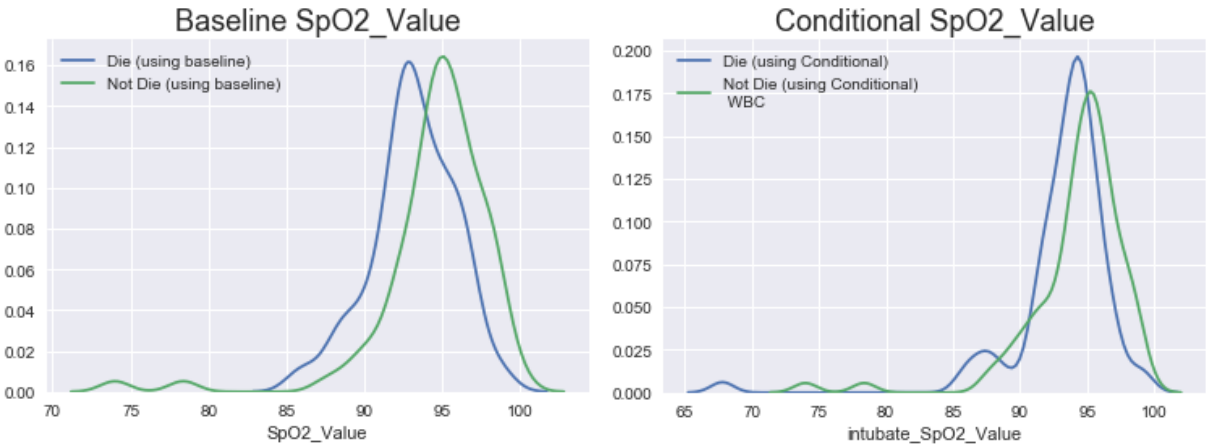
| **Participants** | **Prediction for Ventilated patients # 159 (who have both baseline and conditional features)** | |
|---|---|---|
| Output classes | Recovery vs (Decease) | Recovery vs (Decease) |
| AUC | 0.763 | 0.617 |
| Confusion Matrix<br>Cutoff = 0.5 | [[58. 24.]<br>[24. 53.]] | [[52. 30.]<br>[38. 39.]] |
| PPV | 0.688 | 0.565 |
| NPV | 0.707 | 0.577 |
| Sensitivity | 0.688 | 0.506 |
| specificity | 0.707 | 0.634 |

To further investigate the superiority of baseline features, we plot in Figure 11 the death probability predicted by logistic regression. Using baseline features, the logistic regression tends to predict a distinguishably higher death probability for those who indeed decease at the end. Regarding the distribution of underlying features, the distribution of $SpO_2$ is presented in Figure 12. The distinct laboratory measurements (between survivors and non-survivors) contribute to a high prognostic performance using baseline features.




(a) (b)

**Figure 11 Predicted death probability on ventilated patients: (a) using baseline features, (b) using conditional features.**




(a) (b)

**Figure 12 Distributions of: (a) baseline $SpO_2$ value, (b) conditional $SpO_2$ value for ventilated patients grouping by their final outcomes.**

Our results indicate that the baseline measurements provide a high predictive value. For instance, when predicting mortality using baseline features resulted in a more accurate prediction as opposed to using the features collected during the time of ventilation. We observed that when the patients are ventilated their lab features have universally deteriorated. We surmise that the

patients with extremely abnormal baseline lab measures have already endured severe disease manifestation before the medical intervention, and therefore they are more likely to suffer from irreversible organ damage. As a result, these patients present with a higher risk of mortality. Besides emphasizing the predictive values of baseline features, this finding advocates early hospitalization before symptoms worsen.

Future work with this model includes implementing an early warning system for human-in-the-loop decision making [69]. With the insights gleaned from emerging clinical data, the use of optimized prone positions, medical therapy with antivirals, and anti-inflammatory medication may alleviate the inflammatory response, improve oxygenation, reduce the risk of intubation, and reduce mortality in patients with COVID-19. An approach like the one introduced here can also identify patients for whom early discharge is safe. A triage tool for sorting high- vs. low-risk individuals with COVID-19 would be highly useful in resource-constrained situations in which bed capacity must be tightly managed. Furthermore, our analyses did not include radiological data, which can provide further information regarding the actual extent of the disease. We will investigate the potential of including radiological and natural language features in the prognostic model. Last but not the least, there is a potential to apply multi-task learning techniques [70] to jointly learn at different hospitals, so as to enhance the generality of learned models and incorporate more data into the training process.

# REFERENCES


[1] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *N. Engl. J. Med.*, 2020, doi: 10.1056/NEJMoa2001017.

[2] T. A. Ghebreyesus, "WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020," World Health Orgnatization, 2020.

[3] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study," *Lancet*, vol. 395, no. 10229, pp. 1054–1062, 2020, doi: 10.1016/S0140-6736(20)30566-3.

[4] X. Zhang *et al.*, "Viral and host factors related to the clinical outcome of COVID-19," *Nature*, 2020, doi: 10.1038/s41586-020-2355-0.

[5] A. Yang *et al.*, "Clinical and epidemiological characteristics of COVID-19 patients in Chongqing China," *Front. Public Heal.*, vol. 8, no. May, pp. 1–8, 2020, doi: 10.3389/fpubh.2020.00244.

[6] P. Goyal *et al.*, "Clinical characteristics of COVID-19 in New York City," *N. Engl. J. Med.*, 2020, doi: 10.1056/NEJMc2010419.

[7] S. Richardson *et al.*, "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area," *JAMA - J. Am. Med. Assoc.*, 2020, doi: 10.1001/jama.2020.6775.

[8] P. Ghelichkhani and M. Esmaeili, "Prone position in management of COVID-19 patients; A commentary," *Arch. Acad. Emerg. Med.*, 2020, doi: 10.22037/aaem.v8i1.674.

[9] C. Guérin *et al.*, "Prone positioning in severe acute respiratory distress syndrome," *N. Engl. J. Med.*, 2013, doi: 10.1056/NEJMoa1214103.

[10] A. E. Thompson, B. L. Ranard, Y. Wei, and S. Jelic, "Prone positioning in awake, nonintubated patients with COVID-19 hypoxemic respiratory failure," *JAMA Intern. Med.*, 2020, doi: 10.1001/jamainternmed.2020.3030.

[11] J. H. Beigel *et al.*, "Remdesivir for the treatment of Covid-19 — Preliminary report," *N. Engl. J. Med.*, 2020, doi: 10.1056/nejmoa2007764.

[12] "Dexamethasone in hospitalized patients with Covid-19 — Preliminary report," *N. Engl. J. Med.*, 2020, doi: 10.1056/nejmoa2021436.

[13] L. Zeng *et al.*, "Risk assessment of progression to severe conditions for patients with COVID-19 pneumonia: A single-center retrospective study," *medRxiv*, p. 2020.03.25.20043166, Jan. 2020, doi: 10.1101/2020.03.25.20043166.

[14] H. Huang *et al.*, "Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: A retrospective analysis," *medRxiv*, p. 2020.03.28.20045989, Jan. 2020, doi: 10.1101/2020.03.28.20045989.

[15] H. Chen *et al.*, "A retrospective longitudinal study of COVID-19 as seen by a large urban hospital in Chicago," *medRxiv*, 2020, doi: 10.1101/2020.11.29.20240606.

[16] J. Sarkar and P. Chakrabarti, “A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19,” *medRxiv*, p. 2020.03.25.20043331, Jan. 2020, doi: 10.1101/2020.03.25.20043331.

[17] J. Gong *et al.*, “A tool to early predict severe 2019-novel coronavirus pneumonia (COVID-19) : A multicenter study using the risk nomogram in Wuhan and Guangdong, China,” *medRxiv*, p. 2020.03.17.20037515, Jan. 2020, doi: 10.1101/2020.03.17.20037515.

[18] Y. Shi, X. Yu, H. Zhao, H. Wang, R. Zhao, and J. Sheng, “Host susceptibility to severe COVID-19 and establishment of a host risk score: Findings of 487 cases outside Wuhan,” *Crit. Care*, 2020, doi: 10.1186/s13054-020-2833-7.

[19] United States Census Bureau, “QuickFacts: Los Angeles city, California; New York city, New York; Chicago city, Illinois.” [Online]. Available: https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia,newyorkcitynewyork,chicagocityillinois/PST045219.

[20] World Population Review, “US cities by population.” [Online]. Available: https://worldpopulationreview.com/us-cities.

[21] K. M. J. Azar *et al.*, “Disparities in outcomes among COVID-19 patients in a large health care system in California,” *Health Aff.*, 2020, doi: 10.1377/hlthaff.2020.00598.

[22] M. Hendryx and J. Luo, “COVID-19 prevalence and mortality rates in association with Black race and segregation in the United States April 1 to April 15, 2020,” *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3582857.

[23] L. Holmes *et al.*, “Black–white risk differentials in covid-19 (Sars-cov2) transmission, mortality and case fatality in the united states: Translational epidemiologic perspective and challenges,” *Int. J. Environ. Res. Public Health*, 2020, doi: 10.3390/ijerph17124322.

[24] D. R. Holtgrave, M. A. Barranco, J. M. Tesoriero, D. S. Blog, and E. S. Rosenberg, “Assessing racial and ethnic disparities using a COVID-19 outcomes continuum for New York State,” *Ann. Epidemiol.*, 2020, doi: 10.1016/j.annepidem.2020.06.010.

[25] N. P. Joseph *et al.*, “Racial and ethnic disparities in disease severity on admission chest radiographs among patients admitted with confirmed COVID-19: A retrospective cohort study,” *Radiology*, 2020, doi: 10.1148/radiol.2020202602.

[26] E. G. Price-Haywood, E. G. Price-Haywood, J. Burton, D. Fort, and L. Seoane, “Hospitalization and mortality among black patients and white patients with Covid-19,” *N. Engl. J. Med.*, 2020, doi: 10.1056/NEJMsa2011686.

[27] M. A. Raifman and J. R. Raifman, “Disparities in the population at risk of severe illness from COVID-19 by race/ethnicity and income,” *Am. J. Prev. Med.*, 2020, doi: 10.1016/j.amepre.2020.04.003.

[28] C. T. Rentsch *et al.*, “Covid-19 by race and ethnicity: A national cohort study of 6 million United States veterans,” *medRxiv*, 2020, doi: 10.1101/2020.05.12.20099135.

[29] T. Selden and T. Berdahl, “COVID-19 and racial/ethnic disparities in health risk, employment, and household composition: Study examines potential explanations for racial-ethnic disparities in COVID-19 hospitalizations and mortality,” *Health Aff.*, 2020, doi: 10.1377/hlthaff.2020.00897.

[30] M. J. Townsend, T. K. Kyle, and F. C. Stanford, “Outcomes of COVID-19: Disparities in obesity and by ethnicity/race,” *Int. J. Obes. Suppl.*, 2020, doi: 10.1038/s41366-020-0635-2.

[31] U. of Oxford, “Statistics and research coronavirus (COVID-19) vaccinations.” [Online]. Available: https://ourworldindata.org/covid-vaccinations.

[32] P. Olliaro, “What does 95% COVID-19 vaccine efficacy really mean?” *Lancet Infect. Dis.*, vol. 21, no. 6, p. 769, 2021.

[33] G. Iacobucci, “Covid-19: Single vaccine dose is 33% effective against variant from India, data show,” British Medical Journal Publishing Group, 2021.

[34] N. Phillips, “The coronavirus is here to stay-Here’s what that means,” *Nature*, vol. 590, no. 7846, pp. 382–384, 2021.

[35] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning-based prediction models in healthcare,” *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, 2020, doi: 10.1002/widm.1379.

[36] P. H. Tsai *et al.*, “Clinical manifestation and disease progression in COVID-19 infection,” *Journal of the Chinese Medical Association*. 2021, doi: 10.1097/JCMA.0000000000000463.

[37] H. K. Siddiqi and M. R. Mehra, “COVID-19 illness in native and immunosuppressed states: A clinical-therapeutic staging proposal,” *J. Hear. Lung Transplant.*, 39(5), 2020, doi: 10.1016/j.healun.2020.03.012.

[38] A. Mody *et al.*, “The clinical course of coronavirus disease 2019 in a US hospital system: A multistate analysis,” *Am. J. Epidemiol.*, 2020, doi: 10.1093/aje/kwaa286.

[39] J. Gong *et al.*, “A tool to early predict severe corona virus disease 2019 (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China,” *medRxiv*. 2020, doi: 10.1101/2020.03.17.20037515.

[40] H. C. Acar *et al.*, “An easy-to-use nomogram for predicting in-hospital mortality risk in COVID-19: A retrospective cohort study in a university hospital,” *BMC Infect. Dis.*, 2021, doi: 10.1186/s12879-021-05845-x.

[41] Y. Zhou *et al.*, “Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: A multi-center study in China,” *Scand. J. Trauma. Resusc. Emerg. Med.*, 2020, doi: 10.1186/s13049-020-00795-w.

[42] X. Jiang *et al.*, “Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity,” *Comput. Mater. Contin.*, 2020, doi: 10.32604/cmc.2020.010691.

[43] W. Liang *et al.*, “Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19,” *JAMA Intern. Med.*, 2020, doi: 10.1001/jamainternmed.2020.2033.

[44] J. Xie *et al.*, “Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19,” *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3562456.

[45] C. Iwendi *et al.*, “COVID-19 patient health prediction using boosted random forest algorithm,” *Front. Public Heal.*, 2020, doi: 10.3389/fpubh.2020.00357.

[46] L. Yan *et al.*, “An interpretable mortality prediction model for COVID-19 patients,” *Nat. Mach. Intell.*, 2020, doi: 10.1038/s42256-020-0180-7.

[47] L. Meng *et al.*, “A deep learning prognosis model help alert for COVID-19 patients at high-risk of death: A multi-center study,” *IEEE J. Biomed. Heal. Informatics*, 2020, doi: 10.1109/JBHI.2020.3034296.

[48] J. Lee, J. H. Kim, C. Ta, C. Liu, and C. Weng, “Severity prediction for COVID-19 patients via recurrent neural networks,” *medRxiv*. 2020, doi: 10.1101/2020.08.28.20184200.

[49] M. Fakhfakh, B. Bouaziz, F. Gargouri, and L. Chaari, “ProgNet: COVID-19 prognosis using recurrent and convolutional neural networks,” *Open Med. Imaging J.*, 2021, doi: 10.2174/1874347102012010011.

[50] Wikipedia contributors, “Nomogram,” *Wikipedia, The Free Encyclopedia*. 2021.

[51] L. Glasser and R. Doerfler, “A brief introduction to nomography: Graphical representation of mathematical relationships,” *Int. J. Math. Educ. Sci. Technol.*, vol. 50, no. 8, pp. 1273–1284, 2019.

[52] X. Jiang *et al.*, “Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity,” *Comput. Mater. Contin.*, vol. 62, no. 3, pp. 537–551, 2020, doi: 10.32604/cmc.2020.010691.

[53] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[54] P. Nisha, U. Pawar, and R. O’Reilly, “Interpretable machine learning models for assisting clinicians in the analysis of physiological data,” in *CEUR Workshop Proceedings*, 2019.

[55] J. Chen *et al.*, “COVID-19 infection: the China and Italy perspectives,” *Cell Death and Disease*. 2020, doi: 10.1038/s41419-020-2603-0.

[56] J. W. Frazier, “Race, ethnicity, and place in a changing America: A perspective,” in *Race, Ethnicity, and Place in a Changing America*, 2010.

[57] J. Irvin *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 590–597.

[58] X. Li *et al.*, “Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in COVID-19 patients: A systematic review and meta-analysis,” *Crit. Care*, vol. 24, no. 1, pp. 1–10, 2020.

[59] Y. Liu *et al.*, “Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19,” *J. Infect.*, 2020.

[60] W. Huang *et al.*, “Decreased serum albumin level indicates poor prognosis of COVID-19 patients: hepatic injury analysis from 2,623 hospitalized cases,” *Sci. China Life Sci.*, vol. 63, no. 11, pp. 1678–1687, 2020.

[61] J. Wu, L. Shi, P. Zhang, Y. Wang, and H. Yang, “Is creatinine an independent risk factor for predicting adverse outcomes in COVID-19 patients?” *Transpl. Infect. Dis.*, p. e13539, 2020.

[62] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, “Developing prediction models for clinical use using logistic regression: An overview,” *J. Thorac. Dis.* 2019, doi: 10.21037/jtd.2019.01.25.

[63] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Int. Jt. Conf. Artif. Intell.*, 1995.

[64] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[65] J. Friedman, T. Hastie, and R. Tibshirani, “glmnet: Lasso and elastic-net regularized generalized linear models,” *R Packag. version*, 2009.

[66] E. R. Ziegel, “The elements of statistical learning,” *Technometrics*, 2003, doi: 10.1198/tech.2003.s770.

[67] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Trans. Syst. Man Cybern.*, 1991, doi: 10.1109/21.97458.

[68] R. I. Lerman and S. Yitzhaki, “A note on the calculation and interpretation of the Gini index,” *Econ. Lett.*, 1984, doi: 10.1016/0165-1765(84)90126-5.

[69] Y. Varatharajah, H. Chen, A. Trotter, and R. Iyer, “A dynamic human-in-the-loop recommender system for evidence-based clinical staging of COVID-19,” in *CEUR Workshop Proceedings*, 2020, vol. 2684, pp. 21–22.

[70] H. Chen, Y. Yang, and C. Shao, “Multi-task learning for data-efficient spatiotemporal modeling of tool surface progression in ultrasonic metal welding,” *J. Manuf. Syst.*, vol. 58, pp. 306–315, 2021, doi: https://doi.org/10.1016/j.jmsy.2020.12.009.