IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 18, 2022, accepted April 9, 2022, date of publication April 18, 2022, date of current version April 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3168045

Prediction of Length of Stay in the Emergency Department for COVID-19 Patients: A Machine Learning Approach

EGBE-ETU ETU[1,2], LESLIE MONPLAISIR[1], SUZAN ARSLANTURK[3], SARA MASOUD[1], (Member, IEEE), CELESTINE AGUWA[1], IHOR MARKEVYCH[4], AND JOSEPH MILLER[5]

[1]Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48202, USA
[2]Department of Marketing and Business Analytics, San Jose State University, San Jose, CA 95192, USA
[3]Department of Computer Science, Wayne State University, Detroit, MI 48202, USA
[4]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[5]Departments of Emergency Medicine and Internal Medicine, Henry Ford Hospital, Detroit, MI 48202, USA

Corresponding author: Egbe-Etu Etu (egbe-etu.etu@sjsu.edu)

This work was supported in part by the Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Henry Ford Health System Institutional Review board under Application No. 14426.

ABSTRACT The coronavirus disease (COVID-19) outbreak has become a global public health threat. The influx of COVID-19 patients has prolonged the length of stay (LOS) in the emergency department (ED) in the United States. Our objective is to develop a reliable prediction model for COVID-19 patient ED LOS and identify clinical factors, such as age and comorbidities, associated with LOS within a ''4-hour target.'' Data were collected from an urban, demographically diverse hospital in Detroit for all COVID-19 patients' ED presentations from March 16 to December 29, 2020. We trained four machine learning models, namely logistic regression (LR), gradient boosting (GB), decision tree (DT), and random forest (RF), across different data processing stages to predict COVID-19 patients with an ED LOS of less than or greater than 4 hours. The analysis is inclusive of 3,301 COVID-19 patients with known ED LOS, and 16 significant clinical factors were incorporated. The GB model outperformed the baseline classifier (LR) and tree-based classifiers (DT and RF) with an accuracy of 85% and F1-score of 0.88 for predicting ED LOS in the testing data. No significant accuracy gains were achieved through further splitting. This study identified key independent factors from a combination of patient demographics, comorbidities, and ED operational data that predicted ED stay in patients with prolonged COVID-19. The prediction framework can serve as a decision-support tool to improve ED and hospital resource planning and inform patients about better ED LOS estimations.

INDEX TERMS COVID-19, length of stay (LOS), 4-hour target, emergency department (ED), machine learning.

I. INTRODUCTION

The coronavirus (COVID-19) pandemic has strained healthcare systems in the United States and the world through increased care complexity, the need for medical staff and patient safety, and surges in patients suspected or infected with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). Hospital emergency departments (EDs) face challenges, as the influx of infected COVID-19 patients has strained existing resources.

Due to the pandemic, numerous health systems in the US have reported increased workload and surges in patient volumes, resulting in ED crowding, which harms patient outcomes and puts additional strain on medical staff [1]–[3]. A key characteristic of crowding is the formation of queues in various parts of the health system as a result

The associate editor coordinating the review of this manuscript and approving it for publication was Santosh Kumar.

of demand-exceeding capacities. These queue formations usually lead to extended average ED length of stay (LOS) [4], [5]. A prolonged ED LOS is associated with higher morbidity and mortality [6]–[8]. Numerous health systems have set time-based targets, requiring patients leaving the ED within the first 4 hours of arrival (i.e., ''4-hour target'') [9]–[12]. However, with the ongoing pandemic, this 4-hour target has been hard to reach for COVID-19 patients leading to overcrowding, operational inefficiencies, and higher utilization of hospital resources.

Previous studies conducted before the COVID-19 pandemic on factors associated with ED LOS employed models such as multiple linear regression, logistic regression, decision trees, and accelerated failure time models [13]–[15]. Machine learning techniques can consider a larger number of attributes (i.e., patient records and hospital information) and permutations, which have the potential to yield a better understanding of complex problems and identify factors to predict COVID-19 ED patients' LOS. To the best of our knowledge, no study has combined these data (i.e., patient and ED operational data) to predict COVID-19 ED patients' LOS. In the present study, we innovatively applied four machine learning techniques, namely logistic regression, gradient boosting, decision tree, and random forest algorithm, across different data processing stages to develop a model to accurately predict the ED LOS of COVID-19 patients.

This research aims to first identify risk factors (e.g., age, race, sex, comorbidities, and nurse/physician availability) by examining both patients' medical records and ED operational data, and then predict the ED LOS of COVID-19 patients with respect to a 4-hour target and identify patients with prolonged ED LOS accordingly. Such a model would enable ED operational leadership to understand the factors for prolonged ED LOS and potentially develop targeted interventions to reduce the proportion of COVID-19 patients experiencing long stays, leading to improved resource planning. Furthermore, the prediction model may be useful as a real-time decision support tool to improve care coordination, while also providing feedback to patients on their LOS.

II. METHODS

A. STUDY DESIGN, POPULATION, SETTING, AND OUTCOME

This is a retrospective and prognostic study on the performance of a prediction model. The goal of this study is to develop a prediction model to predict the ED LOS of COVID-19 patients (i.e., LOS <4 hours or ≥ 4 hours). Patient-level data were retrospectively retrieved from the data warehouse of the Henry Ford Hospital electronic health records (EHR) from March 16, 2020, to December 29, 2020. The research team accessed the hospital database/records on April 26, 2021, to obtain the retrospective data used in this study. The Henry Ford Hospital Institutional Review Board approved the study and waived the requirement for informed consent.

Henry Ford Hospital (HFH) is an 877-bed urban academic hospital and research center located in Detroit, Wayne County, Michigan, with an approximate population of 1.75 million. HFH is designated as a Level 1 trauma center that serves a high-acuity and racially diverse urban patient population. The ED treats an estimated 100,000 patients annually (45%). Patients requiring admission must be reviewed in the ED by a specialty-receiving doctor before they move to the ward. According to the John Hopkins Coronavirus Resource Center, COVID-19 cases in Wayne County are among the highest in the country [16].

The total number of ED visits during the study period was 57,665. We included patients with a positive real-time polymerase chain reaction (RT-PCR) result for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) during or before their hospital encounter. We excluded patients who visited the ED during the study period and did not have RT-PCR confirmation of COVID-19. We retrieved de-identified COVID-19 patient-level data, including demographic information, comorbidities, common symptoms, vital signs, and clinical outcomes, from the hospital's EHR for analysis in addition to ED operations data for the given period.

The study sample included adults, of whom 74.7% were African Americans, 5.5% were White, 5.1% were Asians, 0.5% were Hispanics, and 14.2% did not identify their race. Of the sample, 64.9% were female and 13.3% were aged 65 years and above. At least one chronic condition was present in 85.5% of the patients. The demographics of the study cohort mirrored those of the demographic data from the larger Detroit community. The primary study outcome was a length of stay of <4 hours in the ED (i.e., will achieve the ''4-hour target'' or not).

We developed an analytical prediction framework aimed at determining whether the ED LOS for COVID-19 patients will be less than or greater than 4 hours (Figure 1). Figure 1 describes how the data are extracted, processed, and split into training and testing datasets for analysis. We applied four machine learning algorithms for the prediction and conducted a model evaluation to determine the performance of each model. We implemented the four machine learning algorithms on two different data processing stages, first modeling the raw data after feature selection and engineering techniques, and then applying a balancing technique after data cleaning, feature selection, and engineering.

B. DATA COLLECTION AND ANALYSIS

EHR data included 127 clinical and operational variables. Noteworthy variables included age, sex, emergency severity index (ESI), insurance type, comorbid conditions, and patient complaint at the time of presentation. The outcome variable was ED LOS less than or greater than 4 hours and was determined by the first documented contact of the patient with the ED and the time that the patient was documented leaving the ED.

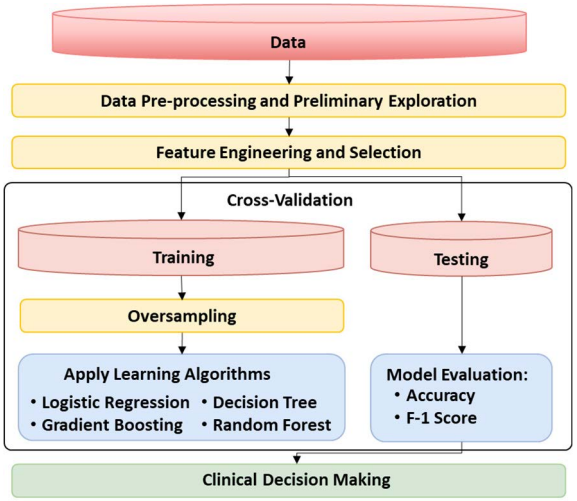


FIGURE 1. Prediction framework for COVID-19 patient ED length of stay. Data extracted from the hospital database is pre-processed and cleaned for analysis. Models are developed to predict COVID-19 patients ED LOS.

1) DATA PRE-PROCESSING AND PRELIMINARY EXPLORATION

We excluded patient features from the analysis that were not available when the patient first arrived at the ED. These features include laboratory tests or other testing results. We imputed the missing data points for each remaining clinical variable with a mean value. We conducted descriptive statistics and data visualizations to check for outliers, distributions, and variations among attributes. The Kolmogorov–Smirnov test (K-S test) was used to analyze data normality. We performed a one-way analysis of variance (ANOVA) to determine whether there were any statistically significant differences between the median ED LOS and patient race. We also conducted a multiple comparison test using the Wilcoxon rank-sum test with continuity correction to show which groups (i.e., patient race) differed from each other with respect to ED LOS. The Bonferroni method was used to adjust the *p-values*.

2) FEATURE ENGINEERING AND SELECTION

We first normalized the data and performed a correlation analysis to understand the correlation between the variables for the feature selection process. The normality results of the K-S test will help us identify the correlation method (i.e., Pearson or Spearman correlation) for feature selection. The K-S test yielded $p < 0.05$, indicating that the data significantly deviated from a normal distribution. Therefore, we used Spearman correlation for feature selection. Spearman's correlation test measures the monotonic relationship between the clinical variables of patients with COVID-19. The correlation coefficient values range from $+1$ to -1 and evaluate the degree of correlation between the two variables. We used one-hot encoding to convert the categorical data into binary variables. The length of ED stay was reclassified as binary (i.e., either '<4 hours' or '≥4 hours'). We used a binning method to convert continuous attributes into categorical variables. Sixty (60) features remain after the feature-engineering process.

3) OVERSAMPLING

Imbalanced classification problems occur when the distribution of observations across known classes (e.g., 2,153 cases with LOS ≥4 hours and 1,148 cases with LOS <4 hours) is skewed. Imbalanced classifications are challenging for predictive modeling because most machine learning algorithms are designed based on the assumption that the model has an equal chance of learning for each class [17]. An imbalanced training dataset violates this assumption, leading to the development of models with poor predictive performance, specifically for the minority class (e.g., 1,148 cases with LOS <4 hours). An oversampling technique called the synthetic minority oversampling technique (SMOTE) was employed to address the imbalanced dataset, where observations from the minority class were randomly duplicated. SMOTE generates synthetic samples from the minority class using information available from the given dataset [18]. The addition of these duplicated values to the minority class balanced the training dataset, providing the model with an equal chance of learning. The application of SMOTE to an imbalanced dataset helps improve the performance of machine learning algorithms when compared to models without any data imbalance technique applied. The oversampling technique adjusted the ratio between these two groups to achieve a ratio of 1:1 (2153:2153).

4) PREDICTION MODEL

We applied four learning algorithms for our analysis: logistic regression, gradient boosting, decision tree, and random forest. Logistic regression (LR) is used as the baseline machine learning algorithm in this study. Gradient boosting (GB) is a machine learning algorithm that combines multiple weak learning models to create a solid predictive model. We built a binary GB classifier model to predict ED LOS. The decision tree (DT) algorithm is a commonly used data mining method for developing classification models based on multiple covariates or developing prediction algorithms for a target variable [19], [20]. We used the DT model to build a tree that identifies all possible attribute combinations from the predictive model, and the proportion of COVID-19 patients within the tree experiencing ED LOS less than or greater than 4 hour was calculated. Random forest (RF) is an ensemble classification algorithm consisting of multiple decision trees [21]. Therefore, we developed an RF prediction model to analyze the interactions between patient-level and ED characteristic data to predict the ED LOS of COVID-19 patients. We divided the sample into 80% training data (March 16 – October 29, 2020) and 20% testing data (October 30 – December 29, 2020). The testing dataset served as a temporal validation for the prediction framework. The models were trained using the introduced datasets. To validate the model, 10-fold cross-validation was used, which required a minimum of two records per leaf with a confidence

factor of 0.3. It is important to note that the developed machine learning models were compared across different data processing stages. Statistical analysis (i.e., t-test) was conducted to evaluate whether significant improvement(s) existed in the models within the two data-processing stages (e.g., comparing RF1 and RF2).

5) MODEL EVALUATION

The performance of each model was evaluated based on accuracy and the F-1 score. The F-1 score is the harmonic mean of precision (positive predictive value) and recall (sensitivity), which complements the accuracy metric [22]. The F-1 score reached its best value at 100% and was worse at 0%. Because we are dealing with an imbalanced dataset in this study (i.e., 2,153 cases with LOS $\geq$4 hour and 1,148 cases with LOS <4 hour), reporting the F-1 score provides a better explanation of the reported accuracy of our prediction models. We also measured the predictive accuracy of the model by reporting the area under the receiver operating characteristic (ROC) curve (AUC). For more information on the model evaluation, see the Appendix. All analyses were conducted using Python (Jupyter version 1.0.0, United States) and IBM SPSS Statistics (version 27.0; 2020, United States).

III. RESULTS

A. PATIENT SELECTION

Figure 2 illustrates a flowchart of the COVID-19 patient cohort. A total of 57,665 patient visits were recorded at the HFH between March 16 and December 29, 2020. A total of 38,189 patients were screened for COVID-19 and 19,476 were not screened. There were 3,301 patients with confirmed COVID-19 diagnoses and 34,888 with non-confirmed COVID-19 diagnoses. The final study cohort used in this analysis was 3,301 patients with confirmed COVID-19 diagnosis (median age, 51 years; 64.9% female), of which 2,153 (65.2%) had an ED LOS $\geq$ of 4 hours.

B. COVID-19 ED PATIENT DATA DESCRIPTION

The COVID-19 patients' demographics and clinical outcomes are displayed in Table 1. We found that 440 patients (13.3%) were older than 65 years, and the median patient age was 51 (IQR 40 – 57) years, showing the prevalence of COVID-19 in older patients. Of the study patients, 35.1% were male and 74.7% were African Americans. The most frequent chief complaint of COVID-19 patients on admission from the ED was shortness of breath (29.7%), followed by fatigue (6.6%), and fever (5.8%). Comorbid conditions were common amongst patients, in particular hypertension (63.9%), obesity (46.4%), anemia (26.1%), type 2 diabetes (17.1%), diabetes (17.0%), congestive heart failure (20.7%) and chronic kidney disease (16.3%). Among the COVID-19 patients included in this study, 808 (24.5%) were admitted to the hospital, 2,144 (65%) were discharged, 122 (3.7%) left without completing service (LWCS), 23 (0.69%) were discharged against medical advice, 173 (5.2%) were placed in observation, and 31 (0.9%) were transferred to another facility.

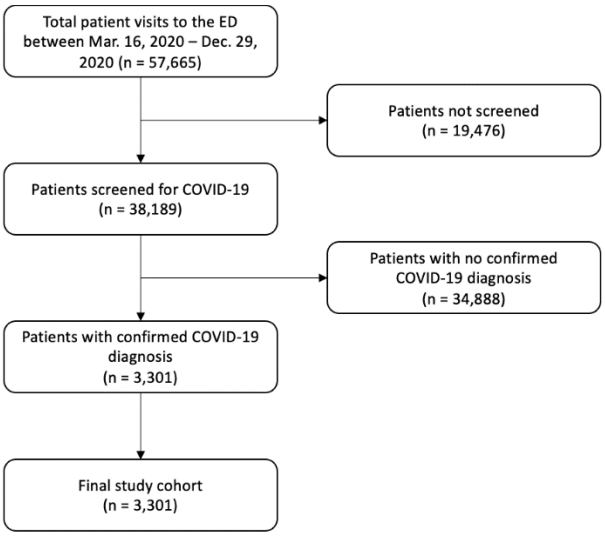


FIGURE 2. Flowchart of COVID-19 patient cohort. Of the 57,665 total ED patient visits within the study period, 3,301 patients with laboratory-confirmed COVID-19 diagnosis who did not meet the exclusion criteria formed the final study cohort.

A statistical visualization of each feature's underlying relationships with the dependent variable (ED LOS) is presented in Figure 3. As displayed in Figure 3(a), the median ED LOS was 5.21 (SD 5.95) hours for COVID-19 patients with a right skewed distribution. Figure 3(b) depicts the variation in COVID-19 patients' ED LOS as age increased, and the median age was 51 (SD 15.0) years. The largest spread was seen for the 60-69 and 70-79 age group, who spent more time in the ED. The test for normality showed that the patient age significantly deviated from the normal distribution ($p < 0.001$). There was a significant difference between patient race and ED LOS ($p < 0.05$). Multiple comparison testing showed a statistically significant difference in the median ED LOS between African American and Asian patients (4.95 vs 6.03 hours, $p = 0.025$), White and Asian (6.70 vs 6.03 hours, $p = 0.002$), as well as between the White and African American patients (6.70 vs 4.95 hours, $p < 0.05$). However, there was no statistically significant difference between Hispanic patients and other patient groups (Figure 3c). Figure 3(d) shows that the most common COVID-19 patients complain, and which symptoms have the longest average ED LOS. On average, COVID-19 patients with complaints related to fever had the most prolonged ED LOS (~9 hours), followed by those with diarrhea (~8.5 hours), and chills (~8 hours). The statistical visualization of each feature's underlying relationships with the dependent variable (ED LOS) is presented in Figure S2-S3.

C. PREDICTION RESULTS

A matrix heatmap showing the correlations between model variables is presented in Figure S1. Variables with either strongly positive or negative correlation coefficients and

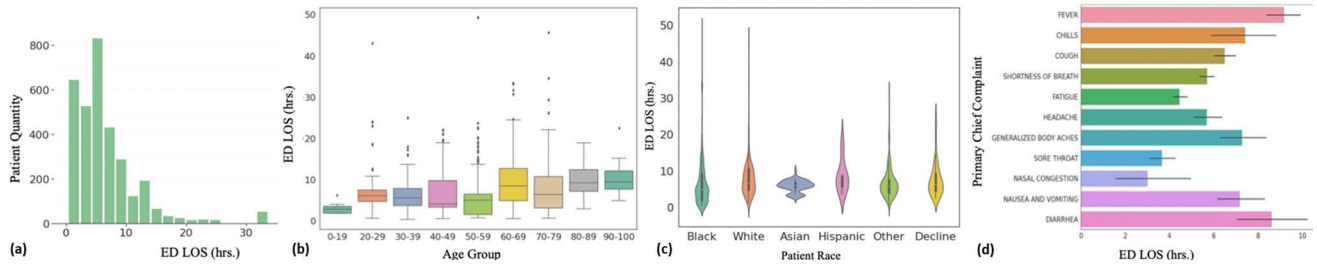


FIGURE 3. COVID-19 patients ED LOS visualization with number of patients, age group, patient race, and primary chief complaint. (a) Histogram of ED LOS, (b) ED LOS variation with age group, (c) ED LOS variation with race, and (d) COVID-19 patient complaint with the most ED LOS.

TABLE 1. Demographic characteristics of COVID-19 patients.

Variable	Number of patients (%)
Number of patients	3,301 (100%)
Race	
Black/African American	2,467 (74.73%)
Asian	167 (5.06%)
White	181 (5.48%)
Hispanic	16 (0.48%)
Others	470 (14.24%)
Age, median (IQR), years	51 (40 – 57)
≥65	440 (13.3%)
<65	2,861 (86.7%)
Gender	
Male	1,158 (35.1%)
Female	2,143 (64.9%)
Common COVID-19 symptoms	
Fever	190 (5.8%)
Cough	170 (5.1%)
Chills	17 (0.5%)
Shortness of breath	980 (29.7%)
Headache	68 (2.1%)
Fatigue	218 (6.6%)
Diarrhea	30 (0.9%)
Nasal congestion	34 (1.0%)
Others	1,594 (48.3%)
Comorbidities	
Hypertension (HTN)	2108 (63.9%)
Diabetes mellitus (DM)	562 (17.0%)
Type 2 DM	563 (17.1%)
Sepsis	41 (1.2%)
Chronic Liver disease (CLD)	200 (6.1%)
Heart Attack	19 (0.6%)
Congestive heart failure (CHF)	684 (20.7%)
Chronic kidney disease (CKD)	539 (16.3%)
Peripheral vascular disease (PVD)	86 (2.6%)
Abdominal Aortic Aneurysm	9 (0.3%)
Atrial Fibrillation	166 (5.0)
Anemia	863 (26.1%)
Coronary Artery Disease (CAD)	143 (4.3%)
Breast Cancer	19 (0.6%)
Chronic obstructive pulmonary disease (COPD)	740 (22.4%)
Obesity	1533 (46.4%)
Vital Signs, Median (IQR)	
Temperature, F	98.3 (98 – 98.8)
Heart rate (HR), bpm	94.0 (85 – 104)
Respiratory rate (RR), rate/min	18.0 (18 – 20)
Percent Oxygen Saturation (SpO_2%)	98.0 (97 – 100)
Systolic blood pressure (SBP), mmHg	134.0 (121 – 146)
Diastolic blood pressure (DBP), mmHg	81.0 (72 – 91)

*Abbreviations: *IQR*, interquartile range.

a p-value less than 0.05 were selected as the best subset of features for building our machine learning model. Our

TABLE 2. Summary of Spearman correlation between ED LOS and clinical features.

Clinical variables	Correlation coefficient (r)	p-value
Physician Availability	0.476	<0.001
Bed Availability	0.264	<0.001
Nurse Availability	0.263	<0.001
Chronic Kidney Disease (CKD)	0.240	<0.001
Diabetes Mellitus (DM)	0.195	<0.001
Emergency Severity Index (ESI)	-0.191	<0.001
Patient Race – Black	-0.149	<0.001
Coronary Artery Disease (CAD)	0.143	<0.001
Chronic Lung Disease (CLD)	0.134	<0.001
Cardiovascular Disease (CVD)	-0.122	<0.001
Obesity	-0.121	<0.001
Atrial Fibrillation	0.119	<0.001
Respiratory Rate (RR)	0.117	<0.001
Percent Oxygen Saturation (SpO_2%)	-0.111	<0.001
Peripheral Vascular Disease (PVD)	0.103	<0.001
Temperature	0.100	<0.001

analysis revealed 16 clinical features with strong correlation coefficients (Table 2). The top four features correlated with ED LOS were physician availability ($r = 0.476$, $p < 0.001$), bed availability ($r = 0.264$, $p < 0.001$), nurse availability ($r = 0.263$, $p < 0.001$), and patients with a history of CKD ($r = 0.239$, $p < 0.001$). The additional significantly correlated features used to build the machine learning models are listed in Table 2.

To optimize the hyperparameters for the different algorithms, we performed grid search cross-validation (GridSearchCV) to determine the optimal parameters for training the GB, DT, and RF algorithms (Table S1). The combination of parameters was evaluated using the accuracy score as a performance metric, and the best parameters with the highest accuracy scores were selected to develop our prediction framework. A section of the decision tree is shown in Figure 4.

Figure 4 shows the decision-tree model. At the first branch level, if a COVID-19 patient presents at the ED and a physician is immediately available, the patient will likely spend <4 hours (denoted as class 0) in the ED; otherwise, we move to the next level. At subsequent branching levels,

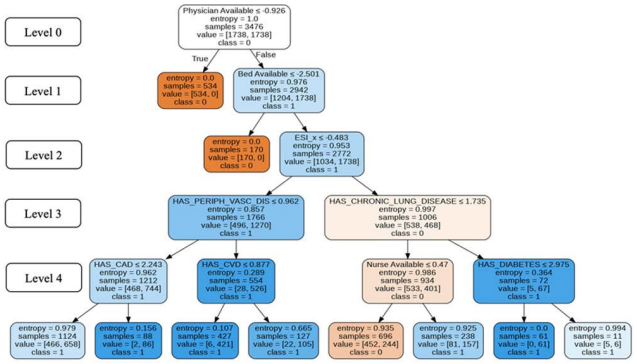


FIGURE 4. A decision tree (DT) on ED operations and COVID-19 patient-level data. The nodes and leaves are presented using rounded rectangles, which represent the outcome of interest. The DT shows two classes, namely class 0 (i.e., patients with ED LOS <4 hours) and class 1 (i.e., patients with ED LOS ≥4 hours). Abbreviation: Emergency Severity Index (ESI), Peripheral Vascular Disease (PVD), Coronary Artery Disease (CAD), Cardiovascular Disease (CVD).

TABLE 3. Temporal validation (i.e., hold-out data) results.

Stage	Model	Precision	Recall	F1-Score	Accuracy	AUC
1	LR	0.76	1.00	0.86	80%	0.88
	GB	0.80	0.95	0.87	82%	0.93
	DT	0.77	1.00	0.87	81%	0.91
	RF	0.85	0.85	0.85	81%	0.90
2	LR	0.76	0.96	0.85	79%	0.88
	GB	0.85	0.92	0.88	85%	0.93
	DT	0.87	0.85	0.86	82%	0.91
	RF	0.92	0.77	0.84	82%	0.92

the logic rule of the decision tree grows to incorporate ED bed availability, ESI, nurse availability, and comorbidities.

Table S2 displays the results of the 10-fold cross validation, whereas Table 3 presents the results of the hold-out data (i.e., temporal validation) of all machine learning algorithms and a comparison between the modeling techniques across different data processing stages.

As observed in Table 3, the first stage (i.e., modeling on raw data + feature engineering + feature selection) shows the performance of LR, GB, DT, and RF on imbalanced data. GB and DT had the highest F1-score (0.87). GB had the highest accuracy (82%) and AUC (0.93) compared to the other regression and tree-based classifiers. The high-performance scores of GB can be attributed to the model's ability to combine multiple weak learning models to create a solid predictive model. In the second stage, we observed a decrease in the F1-score of LR, DT, and RF after implementing feature engineering, feature selection, and oversampling techniques. The GB model performed better in this stage with an F1-score of 0.88 and an accuracy of 85%. In addition, the improved accuracy and AUC scores of the RF model after balancing the data are worth noting. Accuracy is a better metric when working with a balanced dataset, but in reality, most real-world problems are defined around an imbalanced dataset. While accuracy overlooks the false positive and false negatives, the F1-score penalizes any extreme recall or precision.

In this study, the sample size ratio of ≥4 hours and <4 hours ED LOS was 1.87, which is equivalent to 65.22% in favor of ≥4 hours LOS (baseline). Comparing the accuracies reported by all models in Stage 1 (Table 3) to the baseline, we can see that all models significantly improved beyond the baseline. The same notion applies to the second stage, as all models show a better accuracy than the baseline of 50% (i.e., the balanced dataset). Although GB provides the best performance in Stage 2, the main objective of this study is not to promote any specific model, but to develop a decision support tool (i.e., the prediction framework) that is capable of handling data imbalance and feature engineering. Table S2 shows the variance estimates from the 10-fold cross-validation of the AUC of each model. The models showed stable results with a low standard deviation. The T-test results demonstrate that there are significant differences in model performance (i.e., AUC) among the different stages of DT and RF ($p < 0.05$) (Table S3-S6). These results suggest that Stage 2 improves the performance of the models. Specifically, our results suggest that when data cleaning, feature processing, and balancing techniques are implemented, the performance of machine learning models increase.

Figure 5 shows the AUC curves for the trained models. In Figure 5a and 5b, we observe that GB has an AUC of 0.93, which means that it has a good separability measure between the two ED LOS classes. It is important to note that GB has the highest AUC scores for both data processing stages. Next, we observe from the plot that the macro-average AUC curve for RF improves from 0.90 in Stage 1 to 0.92 in Stage 2. Therefore, the RF model performed well in classifying the positive class in the dataset after implementing feature engineering, feature selection, and oversampling techniques. The optimal decision threshold value (i.e., the operational point on the AUC plot) was achieved using Youden's J index [23], which maximizes the difference between the true and false positive values. Table S7 in the Appendix presents the individual thresholds for each model at each stage. The threshold is the point in the AUC plot in which when it is decreased, we obtain more positive values, thus increasing the sensitivity and reducing the specificity of the models. Similarly, when the threshold was decreased, we obtained more negative values, thus reducing the sensitivity, and increasing the specificity. These thresholds can be used when making future probability predictions, which must then be converted from probabilities to crisp class labels (i.e., ED LOS <4 or ≥4 hours).

IV. DISCUSSION

According to the World Health Organization, the COVID-19 virus continues to spread, with estimated global confirmed infection cases of 465.7 million and 6.06 million deaths as of March 17, 2022. According to the US Center for Disease Control and Prevention, as of March 17, 2022, the United States is one of the leading countries with confirmed infection cases of 79.6 million and 970K deaths. During the pandemic,

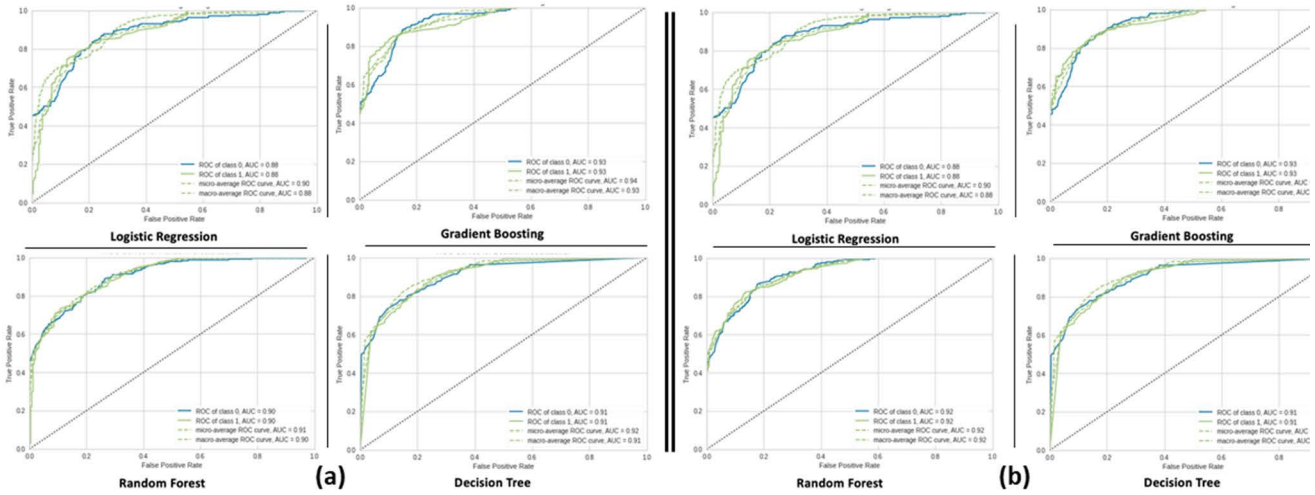


FIGURE 5. ROC curves for the trained models – (a) Stage 1 and (b) Stage 2. It is evident from the plot that the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) is different for Random Forest.

hospitals had to turn away patients due to insufficient oxygen supply and other capacity issues [24]. COVID-19 is becoming increasingly contagious with new variants found, and a prolonged patient ED LOS may lead to higher mortality rates, hospital resource shortages, and the inability to receive new patients [24]. There is a need to predict patient ED LOS to improve planning for limited hospital resources and to develop a decision support system that can improve patient expectations of their LOS.

In the present study, we developed a prediction framework including four models, namely LR, GB, DT, and RF, to predict whether a COVID-19 patient will stay in the ED for less than 4 hours or more. Although both GB and DT were able to achieve superior performance over the other methods, we achieved the best performance using GB when the data were adequately engineered and preprocessed to address the class imbalance issue. It is important to note that most machine-learning models are not capable of handling the class imbalance present in a given dataset. In our dataset, 65.2% of the ED LOS was for patients who stayed for >4 hours in the ED. During prediction, machine learning models trained on imbalanced datasets often favor the majority class, leading to better performance in the majority class than in the minority class [25]. According to [25], [26], developing models that are insensitive to class distribution or utilizing oversampling techniques (e.g., SMOTE) will help address the class imbalance present in a given dataset. Our prediction framework (Stage 2), which combines feature engineering, feature selection, and SMOTE to address the imbalanced dataset, shows that the models can predict the ED LOS of COVID-19 patients, as presented in Table 3. The results showed different performance metrics for the algorithms.

We specifically used the F1-score to measure the performance of the models, as it is a better metric for imbalanced datasets. Focusing on Stage 2, we observed that GB performs better than LR, DT, and RF after engineering the features. One reason, in our opinion, is the ability of the GB classifiers to combine multiple weak learners to create an ensemble of strong classification models. Although GB models are computationally expensive to tune because of the increased number of hyperparameters and sequential nature of the models, their strength lies in the reduction of loss to achieve a more accurate estimate of the response variable. We were careful not to overestimate the complexity of the model and tried to avoid overfitting (i.e., a decrease in prediction performance) when tuning the models. In the literature, RF and GB models have provided varying performance results in different applications, including bioinformatics [27], medical informatics [28], and other domains [29]. To the best of our knowledge, this is the first study to use a machine learning approach to predict COVID-19 patients' ED LOS that combines patient and ED level data. We identified significant factors present upon patient arrival to the ED, such as physician availability, bed availability, nurse availability, ESI, patient race, vital signs, and comorbid conditions. Although these features apply to COVID-19 patients, they have the potential to be tested in future disaster outbreaks unrelated to COVID-19.

Age has previously been demonstrated to contribute significantly to ED LOS [30] but was not a significant factor in our study. Previous studies also identified diverse and heterogeneous factors, such as the time of day that the patient presents, waiting for the emergency physician for greater than two hours, patient sex, and race to contribute to ED LOS [13], [31], [32]. Our research agrees with these findings that physician availability and race contribute to prolonged ED stays for COVID-19 patients. The application of a data-driven machine learning approach has helped to identify significant predictors for solving LOS questions during a pandemic. Applied in real time, our machine learning model could help identify COVID-19 patients more likely to stay longer in the ED during the first moments of their presentation. These machine learning models could be expanded to other outcomes of interest, such as hospital LOS, personalized risk prediction models, and other clinical outcomes.

Prior studies have demonstrated the potential of similar techniques to be applied to non-COVID-19 patients. A study by Gill *et al.* used gradient boosting among low-acuity fast-track patients to predict ED LOS $\geq$4 hours with an AUC of 0.89 [33], although an F1-score that is appropriate when data imbalance exists has not been reported [34]. A data mining technique was utilized to predict ED LOS $\geq$4 hours in a regional Australian public hospital pre-COVID-19 [13]. They reported an accuracy of 85% but did not report an F1-Score [13]. The advantages of our study include the focus on COVID-19, the development of a predictive framework, and evaluation of the framework's performance using accuracy, recall, precision, F1-score, and AUC to ensure proper reporting of analysis performed on imbalanced datasets. In the present study, machine learning models were applied to all confirmed COVID-19 patients present in the ED and incorporated attributes related to patient complaints, medical history, and initial ED characteristics. Stage 2, which combines feature engineering, feature selection, and an oversampling technique to address the imbalanced dataset, shows that the models are capable of predicting ED LOS, as presented in Table 3. In addition, Stage 2 shows a slight performance improvement compared with Stage 1 (i.e., when no balancing technique is applied).

Our study had some limitations. First, some clinical variables had missing values due to administrative errors. Missing data points were imputed with the corresponding mean value, introducing bias if the missingness was not random. Second, our model was limited to features readily available in the hospital's EHR system. Other features that were not captured or retrieved from the EHR, such as hospital capacity during COVID-19 patient presentation or ED boarding, were not included in the study but could impact ED LOS. Third, the identified predictors are significant in the study hospital ED and may not be readily translated to other EDs. With slight modifications to our model, other EDs can use the prediction framework as a tool to identify significant predictors in their data. Lastly, in assessing only COVID-19 patients, our models do not fully account for all ED patients who seek emergency care during a pandemic.

V. CONCLUSION

In conclusion, we describe some of the medical and ED characteristics of COVID-19 patients during hospitalization. The study identified significant factors based on a combination of patient demographics, comorbidities, and ED operational data associated with prolonged stays in COVID-19 patients. We innovatively trained four prediction models on these factors to predict COVID-19 patients' ED LOS. With further validation, the model and results of this study can serve as an effective decision-support tool to improve healthcare delivery/resource planning and help clinicians develop effective interventions to address patient outcomes (e.g., reducing prolonged LOS). Although the models are trained based on locally collected data and clinical information from Henry Ford Hospital, they can be retrained and updated for use in other EDs to predict COVID-19 patient LOS.

ACKNOWLEDGMENT

The authors would like to acknowledge the anonymous reviewers whose feedback helped to improve this manuscript.

REFERENCES

[1] E. Walters, S. Najmabadi, and E. Platoff, "Texas hospitals are running out of drugs, beds, ventilators and even staff," Texas Tribune, Austin, TX, USA, Tech. Rep., 2020.
[2] B. C. Sun, R. Y. Hsia, R. E. Weiss, D. Zingmond, L.-J. Liang, W. Han, H. McCreath, and S. M. Asch, "Effect of emergency department crowding on outcomes of admitted patients," *Ann. Emergency Med.*, vol. 61, no. 6, pp. 605–611, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S019606441201699X
[3] A. Guttmann, M. J. Schull, M. J. Vermeulen, and T. A. Stukel, "Association between waiting times and short term mortality and hospital admission after departure from emergency department: Population based cohort study from Ontario, Canada," *Brit. Med. J.*, vol. 342, p. d2983, Jun. 2011.
[4] U. Hwang, M. L. McCarthy, D. Aronsky, B. Asplin, P. W. Crane, C. K. Craven, S. K. Epstein, C. Fee, D. A. Handel, J. M. Pines, N. K. Rathlev, R. W. Schafermeyer, F. L. Zwemer, Jr., and S. L. Bernstein, "Measures of crowding in the emergency department: A systematic review," *Academic Emergency Med.*, vol. 18, no. 5, pp. 527–538, May 2011.
[5] N. R. Hoot and D. Aronsky, "Systematic review of emergency department crowding: Causes, effects, and solutions," *Ann. Emerg. Med.*, vol. 52, no. 2, pp. 126–136, 2008, doi: 10.1016/j.annemergmed.2008.03.014.
[6] S. Clair, A. Staib, S. Khanna, N. M. Good, J. Boyle, R. Cattell, L. Heiniger, B. R. Griffin, A. J. Bell, J. Lind, and I. A. Scott, "The national emergency access target (NEAT) and the 4-hour rule: Time to review the target," *Med. J. Aust.*, vol. 204, no. 9, p. 354, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.5694/mja15.01177?casa_token=ijKnvDmdfIIAAAAA:8dfpe4v4DN0Yo6d6yB–7f5CIdrUiFV5BZ6yq61odPXxcJRyns33RP1E4G5NmD73cFgjAlDaBsN4NHeq
[7] B. G. Carr, A. J. Kaye, D. J. Wiebe, V. H. Gracias, C. W. Schwab, and P. M. Reilly, "Emergency department length of stay: A major risk factor for pneumonia in intubated blunt trauma patients," *J. Trauma: Injury, Infection Crit. Care*, vol. 63, no. 1, pp. 9–12, 2007.
[8] D. Liew, D. Liew, and M. P. Kennedy, "Emergency department length of stay independently predicts excess inpatient length of stay," *Med. J. Aust.*, vol. 179, no. 10, pp. 524–526, Nov. 2003.
[9] C. Morley, M. Unwin, G. M. Peterson, J. Stankovich, and L. Kinsman, "Emergency department crowding: A systematic review of causes, consequences and solutions," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0203316.
[10] N. Bobrovitz, D. S. Lasserson, and A. D. M. Briggs, "Who breaches the four-hour emergency department wait time target? A retrospective analysis of 374,000 emergency department attendances between 2008 and 2013 at a type 1 emergency department in England," *BMC Emergency Med.*, vol. 17, no. 1, p. 32, Dec. 2017.
[11] P. Jones and K. Schimanski, "The four hour target to reduce emergency department 'waiting time': A systematic review of clinical outcomes," *Emergency Med. Australasia*, vol. 22, no. 5, pp. 391–398, Oct. 2010.
[12] A. Mortimore and S. Cooper, "The '4-hour target': Emergency nurses' views," *Emergency Med. J.*, vol. 24, no. 6, pp. 402–404, 2007.
[13] M. A. Rahman, B. Honan, T. Glanville, P. Hough, and K. Walker, "Using data mining to predict emergency department length of stay greater than 4 hours: Derivation and single-site validation of a decision tree algorithm," *Emergency Med. Australasia*, vol. 32, no. 3, pp. 416–421, Jun. 2020.
[14] C.-H. Chaou, H.-H. Chen, S.-H. Chang, P. Tang, S.-L. Pan, A. M.-F. Yen, and T.-F. Chiu, "Predicting length of stay among patients discharged from the emergency department—Using an accelerated failure time model," *PLoS ONE*, vol. 12, no. 1, Jan. 2017, Art. no. e0165756.
[15] P. Yoon, I. Steiner, and G. Reinhardt, "Analysis of factors influencing length of stay in the emergency department," *Can. J. Emergency Med.*, vol. 5, pp. 155–161, May 2003.
[16] J. Hopkins. *Johns Hopkins Coronavirus Resource Center*. Johns Hopkins University (JHU). Accessed: Mar. 17, 2022. [Online]. Available: https://coronavirus.jhu.edu/us-map
[17] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognit.*, vol. 72, pp. 327–340, Dec. 2017.

[18] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, no. 1, p. 106, 2013, doi: 10.1186/1471-2105-14-106.
[19] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.
[20] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.
[21] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA, USA: Springer, 2012, pp. 307–323.
[22] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.* Berlin, Germany: Springer, 2006, pp. 1015–1021.
[23] A. Zongo, S. Simpson, J. A. Johnson, and D. T. Eurich, "Optimal threshold of adherence to lipid lowering drugs in predicting acute coronary syndrome, stroke, or mortality: A cohort study," *PLoS ONE*, vol. 14, no. 9, Sep. 2019, Art. no. e0223062.
[24] M. Holcombe and L. Mascarenhas, "Colorado identifies first known case of U.K. coronavirus variant in U.S.," CNN, Atlanta, GA, USA, Tech. Rep., 2020. [Online]. Available: https://www.cnn.com/2020/12/29/health/us-coronavirus-tuesday/index.html
[25] C. E. Golden, M. J. Rothrock, and A. Mishra, "Comparison between random forest and gradient boosting machine methods for predicting listeria SPP. Prevalence in the environment of pastured poultry farms," *Food Res. Int.*, vol. 122, pp. 47–55, Aug. 2019.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321–357, Jun. 2002.
[27] J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proc.*, vol. 5, no. S3, pp. 1–5, Dec. 2011, doi: 10.1186/1753-6561-5-S3-S11.
[28] J.-C. Huang, Y.-C. Tsai, P.-Y. Wu, Y.-H. Lien, C.-Y. Chien, C.-F. Kuo, J.-F. Hung, S.-C. Chen, and C.-H. Kuo, "Predictive modeling of blood pressure during hemodialysis: A comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method," *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105536.
[29] S. Nawar and A. M. Mouazen, "Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-NIR spectroscopy measurements of soil total nitrogen and total carbon," *Sensors*, vol. 17, no. 10, p. 2428, 2017.
[30] E. M. Carter and H. W. Potts, "Predicting length of stay from an electronic patient record system: A primary total knee replacement example," *BMC Med. Informat. Decis. Making*, vol. 14, no. 1, p. 26, Dec. 2014.
[31] I. Cheng, D. Taylor, M. J. Schull, M. Zwarenstein, A. Kiss, M. Castren, M. Brommels, M. Yeoh, and F. Kerr, "Comparison of emergency department time performance between a Canadian and an Australian academic tertiary hospital," *Emergency Med. Australasia*, vol. 31, no. 4, pp. 605–611, Aug. 2019.
[32] S. Khanna, J. Boyle, N. Good, and J. Lind, "New emergency department quality measure: From access block to national emergency access target compliance," *Emergency Med. Australasia*, vol. 25, no. 6, pp. 565–572, Dec. 2013.
[33] S. D. Gill, S. E. Lane, M. Sheridan, E. Ellis, D. Smith, and J. Stella, "Why do 'fast track' patients stay more than four hours in the emergency department? An investigation of factors that predict length of stay," *Emergency Med. Australasia*, vol. 30, no. 5, pp. 641–647, Oct. 2018.
[34] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118432.

EGBE-ETU ETU received the Ph.D. degree in industrial engineering from Wayne State University, Detroit, USA.

He is currently an Assistant Professor of business analytics with San Jose State University, San Jose, USA. His research interests include the development of use-inspired machine learning models to solve challenging business problems in healthcare, manufacturing, and transportation.

LESLIE MONPLAISIR received the Ph.D. degree in engineering management from the Missouri University of Science and Technology (MUST), USA.

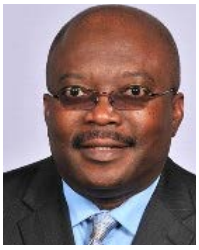
He is currently a Professor of industrial engineering and the Associate Dean of the College of Engineering, Wayne State University, Detroit, USA. His research interests include lean product development, design for lean systems and services and design reuse, new product technology decision modeling, product architecture optimization, and healthcare technology system design.

SUZAN ARSLANTURK received the Ph.D. degree from Oakland University, Michigan.

She is currently an Assistant Professor with the Department of Computer Science, Wayne Wayne State University, Detroit, USA. Her research interests include the broad area of health informatics and healthcare systems engineering. In particular, she is motivated with tackling challenging theoretical and applied research problems that enable developing innovative machine learning and clinical solutions to enhance individual and population health outcomes, improve patient care, and optimize the operational performance of healthcare delivery systems.

SARA MASOUD (Member, IEEE) received the B.S. degree in industrial engineering from the Sharif University of Technology, Tehran, Iran, in 2014, and the M.S. and Ph.D. degrees in statistics and systems and industrial engineering from The University of Arizona, AZ, USA, in 2019.

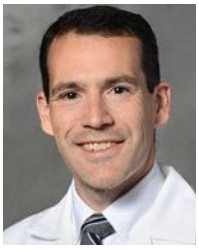
She is currently an Assistant Professor with the Department of Industrial and Systems Engineering, Wayne State University, Detroit, USA. Her research interests include extended reality and dynamic, data-driven application systems by utilizing machine learning, simulation and optimization models in agro-industry, transportation, health care, and manufacturing.



CELESTINE AGUWA received the M.S.I.E. degree in industrial and manufacturing engineering from the University of Pittsburgh, USA, and the Ph.D. degree industrial and manufacturing engineering from the University of Massachusetts, Amherst, USA.

He has a cross-functional industrial experience at Ford Motor Company and extensive professional experience as an Architect. His research interests include lean and value methodology in product development and advanced manufacturing, customer voice analysis, and decision analysis modeling.

IHOR MARKEVYCH is currently pursuing the Master of Computational Data Science degree with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA. His research interests include differential modeling of semi-linear parabolic systems with composition method and convergence of iterations in the trotter-daletskii formula for nonlinear perturbation. He is an Active Member of the Association of Computing Machinery Society.



JOSEPH MILLER is currently an Emergency Medicine Physician with Henry Ford Hospital, Detroit, USA. He specializes in emergency medicine, internal medicine, and clinical research. His research interests include emergency neurological conditions, such as acute stroke, epilepsy, and traumatic brain injury. He also does research in hypertensive emergencies and teaches research methodology within the health systems.