Development and External Validation of the Machine Learning Models to Predict In-Hospital Cardiac Arrest in the Emergency Department: A Cross-Country Approach

Amir Hossein Mostafavi Sterabadi
Mentor: Eric Chou, MD

Abstract

a) Research Question: In Emergency Department (ED) presenting patients within the United States, will our 6 previously internally validated machine-learning (ML) models be able to utilize patient's triage data, vitals, chief complain, and demographics to successfully identify those who have had an emergency department-based cardiac arrest (EDCA) event?

b) Background, Significance, and Rationale
Through our initial approach, we were able to identify utility and predictive strength of ML models for patients at risk of emergency department-based cardiac arrest (EDCA) who presented in an ED in Taiwan. Our cross-country study aims to prove the utility, reliability, and predictive strength of the initial ML models in an ED population within the United States. We hope to provide reliability through an external validation of our initial ML models as a clinical tool to predict and respond appropriately to patients at risk of cardiac arrest who present to the emergency department.

c) Materials and Methods
We utilized the same training cohort models developed from the database of adult patients at a tertiary training hospital in Taiwan between Jan. 1, 2009, to December 31, 2015. We retrospectively collected data from the ED of a tertiary teaching hospital in the United States between August 31, 2019, to December 31, 2020, to be utilized for external validation as the testing cohort of our ML models. In addition, we trained 6 different ML models in the training cohort using patient features such as triage information and clinical symptoms. We then employed K-fold cross validation and evaluated the performance of our models based on the area under the receiver operating characteristic curve (AUC) in the external validation cohort.

d) Results

237,349 and 49,792 patients were included in the training and testing cohort respectively; 477 (0.2%) and 166 (0.3%) were identified to have had an EDCA. All the ML models performed with excellent discrimination based on AUC. Of the constructed ML models, light gradient-boosting machine (LGBM) achieved the best performance of AUC (0.897, 95%, 95% CL: 0.876-0.916) through utility of 7-fold cross validation. There were no significant differences between the constructed models.

e) Conclusions

Through our study we were able to develop and externally validate our constructed ML models for prediction of EDCA in patients presenting to the ED. Our findings suggest that our ML models have the capabilities to be generalized and applicable as a tool to be used in the ED to predict, prevent, and respond to potential EDCA events based on their discriminatory abilities described in the study.

Research Question

In our previous study, the main objective was comparison of the NEWS 2 criteria to 3 supervised ML models in their ability to predict EDCA in a patient population within Taiwan. Our 3 ML models exemplified excellent predictive ability with comparison to the NEWS 2 score in that population.

In our current study our objective is if in Emergency Department (ED) presenting patients within the United States, will our 6 previously internally validated machine-learning (ML) models be able to utilize patient's triage data, vitals, chief complain, and demographics to successfully identify those who have had an emergency department-based cardiac arrest (EDCA) event? We predict that based on our prior research and implementation of similar feature selection, our ML models should show strong predictive power in identifying patients at risk of EDCA events while identifying patients who are not at risk of EDCA events using readily available triage data and features to learn and apply from. Through the goals of our research, our hope is to provide more evidence of the reliability of ML models as a tool in healthcare for improving patient outcomes, resource allocation, and triage efficiency for patients in need of higher level of care for treatment of cardiac arrest events.

Introduction and Significance

In-Hospital Cardiac Arrest

In hospital cardiac arrest (IHCA) is an acute event most commonly defined as loss of circulation requiring prompt resuscitation during a patient's hospitalization.[1] The estimated occurrence of in-hospital arrest in the United States is approximately 290,000 cases annually.[1] Recent recognition of in-hospital cardiac arrest as an essentially different disease from out-of-hospital cardiac arrest (OHCA) has led to independent studies on the disease process.[2,3] In regards to the etiology of the disease process of IHCA, a recent large systematic review and meta-analysis provided evidence of cause of IHCA events with a 95% confidence interval with the most prevalent causes being hypoxia (14.2-38.7%), acute coronary syndrome (13.9%-22.6%), arrhythmias (0-34.9%), infection (9.5-19.3%), and heart failure (6.5-18.8%).[4,5] A recent retrospective multicenter observational study in southern Sweden conducted to identify characteristics and outcomes of patients admitted to the ICU after return of spontaneous circulation (ROSC) identified differences regarding 30-day mortality based on patents IHCA vs. OHCA events.[3] In addition, the study also identified a higher rate of witnessed arrest and shorter delay times in treatment for patients in IHCA vs. OHCA, the difference likely stemming from monitoring and proximity to care.[3] In addition, the research identified that patients who had suffered an IHCA event had a lower 30-day mortality and good long-term neurological outcomes in comparison to patients with OHCA events; there was no noted significant difference in survival curve between the two cohorts.[3] With the discussion of independent disease process between IHCA and OHCA events, there have been limited research regarding the treatment of IHCA with most of the guidelines for management being provided through research on OHCA.[3,6–8] Based on the distinct characteristics of IHCA, there stems a need for more dedicated research not only in the management of IHCA events but also in prediction of these events.

Early Warning Scores (EWS)

Previous research has noted that patients who experience cardiac arrest or requiring ICU admission and management show signs of clinical deterioration several hours before the time of the event.[9–14] The Modified Early Warning Score (MEWS) has been studied as a pre-prognostic

factor in patients who may be at risk of IHCA events[15]. The MEWS scoring system is a utilized system designed to identify patients at risk of deterioration. The scoring system takes into account patient data such as systolic blood pressure, heart rate, respiratory rate, blood pressure and level of consciousness, offering utility as a bedside test to identify patients requiring higher level of medical attention.[15–17] A study on the comparison of the number of IHCA compared between 18 months before and after the introduction of the MEWS system has been reported in a study in Japan which showed a reduction of rate of IHCA events per 1000 admission from 5.21% (79/15,170) to 2.05% (43/17,961).[9] The utility of MEWS as a predictive tool for IHCA events cannot be overlooked but limitations and room for improvement in the scoring system have also been discussed in the research. A previously conducted nested case-control study of 88 patients comparing vital signs and MEWS were compared at admission and 48 hours prior to cardiac arrest (CA).[18] The study found that the MEWS scoring system was significantly different between patients experiencing CA and control patients 48 hours prior to the event, but found that it included poor predictors of CA such as temperature and omitted significant predictors such as diastolic blood pressure (BP) and pulse pressure index.[18] With this in mind, the current utility of MEWS as a predictive factor does take into account evidence based predictors of patients at risk of IHCA events, but leaves room for further studies on predictive modeling inclusive of other significant predictors of IHCA.

In addition to the MEWS and other early warning score (EWS), one scoring system which has shown strong data in its ability to discriminate ward patients at risk of cardiac arrest, death, or ICU admissions is the National Early Warning Score (NEWS).[19] The NEWS was developed by the Royal College of Physicians of London in 2012, taking into account seven parameters (temperature, systolic blood pressure, respiratory rate, oxygen saturation, heart rate, and level of consciousness.[20] A previous retrospective case-control study with the aim of evaluating NEWS in the 24 hour preceding in-hospital cardiac arrest among ward patients noted that there was a 3.17 in odds of IHCA events compared to low-risk category, providing evidence of strong discriminatory properties for patients at risk of these events.[21]

Emergency Department Cardiac Arrest (EDCA)

As mentioned, the focus of predictive modeling is on patients at risk of IHCA who may be identified with signs of clinical deterioration based on EWS systems. Currently, most of the

models have focused on the subset of populations who have been hospitalized or admitted with limited data on predictive modeling of the emergency department patient population. With limited data on this population, the importance of a strong predictive model focused on identifying predictive outcomes is necessary for advancing the level of patient care for those at risk of EDCA events. One model which we highlight is the updated NEWS 2 system. In 2017, an updated version of the NEWS, NEWS 2, was created with additional parameters inclusive of SpO2 scoring scale, and the variable "new confusion" to the alert/verbal/pain/unresponsive score.[22] A Swedish cohort study in 2022 focused on adult cardiac arrest in the emergency department identified within their study population that 1.6 cardiac arrest occur in 10,000 ED visits (10% of total IHCA events) and identified that every fifth patient was not captured by NEWS 2 identifying a need for stronger predictive models for IHCA events of the emergency department presenting patient population.[23]

This brings forth a focus and interest in further studies and evidence based predictive modalities for IHCA events, but also specific studies focused on predictive modeling of IHCA events for patients in the emergency department (EDCA). With the recent advancements and biomedical focus on artificial intelligence, we believe strongly in its utility in expanding on the current available models for prediction of EDCA events.

Machine Learning (ML) Models Overview

Machine learning (ML) is an evolving field, utilizing algorithms designed to emulate human intelligence and learn from large data sets to provide helpful predictions based on provided information.[24] As mentioned, with the growing trend in predictive modeling as well as the utility of artificial intelligence, the field of machine learning has been utilized in a multitude of different fields spanning from finance, entertainment, computer science, as well as medicine.[24] There are many different types of learnings readily utilized by ML models. Supervised learning is a form of machine learning that works to map and infer the function of outputs based on previously prescribed inputs.[25] In this setting, these algorithms are those in which need some external assistance in identifying and pairing connections between the inputs and outputs.[25] This form of external assistance comes from the division of a training and testing data set in which the ML models are provided data in the training set to learn how to match and classify said output

variables.[25] By learning patterns through the training data sets, these ML models can then be applied to a different set of input data and utilize the learned patterns for accurate and efficient predictions.[25] Through supervised learning ML models trained and tested on large data sets, the application and possibilities for their use in the field of medicine remains to be seen. Several previous studies have focused on machine learning to predict cardiac arrest in septic, ward, pediatric, and acute coronary syndrome patients but there are limited studies on machine learning in the predictive ability of EDCA based solely on triage data.[26–30]

Recent Research

As discussed, current research has not only placed emphasis on IHCA but also on the utility of ML models and their ability to predict such events. A retrospective cohort study from 2018 identified utility in the usage of deep learning algorithms to identify patients at risk of IHCA events.[31] In their study focused on admitted patients over 2 hospital systems from 2010 to 2017, they identified that deep learning-based warning system (DEWS) was able to outperform modified early warning score, a random forest algorithm, and logistic regression based on the area under the precision-recall curve (AUPRC).[31] They noted that their DEWS system was able to identify >50% of patients with IHCA events 14-hours prior to the event with low false-alarm rates and high sensitivity with 24% higher sensitivity and 14.6% reduced alarms compared to modified early warning systems.[31] This highlights the ability of artificial intelligence as a strong prediction tool for patients who may be at risk of IHCA events. In addition, this study focused on predictive modeling of IHCA events for patients admitted in the 2 respective hospitals ICU's and general wards.

The focus of our prior and current research is to identify the strength in prediction of our ML models focused on emergency department based patients who may be at risk of EDCA events utilizing solely triage available data, chief complaints, and patient demographics.[26] Over a 7 year period of data from a tertiary teaching hospital in Taiwan, we were able to train and test 3 supervised ML models (Random Forest, Gradient Boosting, and Extra Trees classifier) and compared their predictive ability to that of the NEWS2 scoring system.[26] All of the constructed models showed strong ability to discriminate and identify EDCA based only on the triage information provided in comparison to the NEWS 2 scoring system.[26]

Rationale and Impact

The purpose of our current research is to continue to identify and validate our findings from our previous study to a general population outside of the initial population of ED patients in a tertiary teaching hospital in Taiwan. During our previous research our ML models were able to obtain and represent strong predictive abilities. We hope that through utilization of our models within a cross-country population will provide external validation of our ML models to patient populations presenting to the ED globally.

With limited research on EDCA and continued growth in the field of artificial intelligence we hope to provide a possible ML model to be implemented within EMR systems throughout all EDs to not only identify patients who may be at risk of EDCA events, but also for hospital personnel to be prepared to respond with appropriate resource allocation and utilization for providing the best care. In addition, by providing external validation of our ML models, we hope to continue to provide evidence-based impact of the utility of ML models and their predictive abilities and impact for future studies on predictive model development in the field of medicine.

Materials and Methods

Study Design and Data Collection

Our previous study consisted of a retrospective cohort study utilizing the electronic medical record (EMR) data from the integrated Medical Database (iMD) of The National Taiwan University Hospital (NTUH). During the analysis process of the study, we included a total of 316,465 adult patients above the age of 18; of this population, 237,349 were assigned to the training cohort. The patient population selected encompassed ED visits between January 1, 2009, to December 31, 2015. The initial training cohort of our data was utilized as the cohort for development and training of our current studies ML models. The main focus of our current research was on the prediction of EDCA events which we defined as patients presenting with absence of a palpable pulse despite attempted resuscitation within the ED of BAS that could be identified with the use of our 6 trained ML models.

The current study is also a retrospective cohort study which included the additional data retrieved from the EMR system of Baylor Scott & White All Saints Medical Center (BAS), a tertiary teaching hospital within the United States. The patient population was sampled from patients who presented to the ED between the time of August 31, 2019 and December 31, 2020. Patient data such as demographics, past medical history, ED triage vital signs, as well as laboratory results were retrieved from the EMR system at BAS. All adult patients (18 years of age>) who presented as OHCA or without blood testing were excluded. In addition, patients presenting to ED due to a non-emergent reason (for instance, issuing a certificate) or substantial missing information (loss of vital-signs or loss of age record) were excluded. For patients who had presented to BAS for repeat visits, we selected the last visit per patient to maximize the statistical power of our analysis. After doing so, we then entered data into Microsoft Excel 2010 and came up with further columns and rows for entry of each relevant data point inclusive to the testing model. After the data was entered and completed in the Microsoft excel document, we then reviewed the patient data for missing variables. We replaced the missing variables of data within columns with the mean, median, or mode of the class being evaluated. All adult patients who presented to the BAS ED during the time listed were identified in the study. The study was approved by the Institutional Review Board of NTUH (201606072RINA) and BAS (reference number: 344143) and waived the requirements for informed consent.

After the initial data retrieval, entry, and review, the statistical analysis team at BAS utilized IBM SPSS Statistics for Windows for processing and data analysis testing as discussed below. In efforts to assess normality of the distribution of our data, we employed the Shapiro-wilk test. The Shapiro-wilk test has shown in previous studies to give the most powerful test and the preferred test in most situations of research.[32(p1)] The results of our initial processing on the training data from the initial study were represented by means with standard deviation (SD) or medians with an interquartile range based on the (Shapiro-Wilk) normality test for continuous variables. Percentages were calculated and evaluated for all our categorical variables. For the assessment of our single variable distributions (univariate analysis) within our sample, we utilized student's t-test, Chi-squared test, Fisher's exact test, or Mann-Whitney U test for outcome differences. Overview of the data testing discussed above are listed as follows:

- Student's t-test is a test commonly utilized for continuous data points where the expected values between two groups are the same with the assumption of normality within the distribution of data.[33]
- Fisher's exact test is a test commonly utilized for binary data within unpaired samples.[33]
- Chi-square test is similar test to Fisher's exact test that can be used for comparison of two or more categories of the variable outcomes. Chi-square test has some conditions when utilized for analysis as the sample size of each variable must be above >60 and the expected number of each field above or equal to 5.[33]
- Mann-Whitney test (also known as the Wilcoxon's rank sum test) is a test commonly utilized for continuous data in a sample. In regards to comparison to Student's t-test, Mann-Whitney is able to test data that is not normally distributed and can be used for testing of paired or unpaired data points within a sample.[33]

Variables with a $P<0.1$ on our previous training cohort were chosen as input features to be utilized for the construction of our initial studies ML models. In our previous study a total of 54 clinical features were chosen based on this setting, inclusive of 5 demographics, 35 symptoms, and 14 triage data points based on statistical analysis with a P value of less than 0.1. In our current data we utilized a total of 41 features based on prior constructed models including 2 demographics (age, sex), 9 triage data, and 30 laboratory results.

Machine Learning Models

For our ML models, we utilized 6 supervised ML algorithms for our testing cohort on patients presenting to BAS ED. These models included, Light Gradient Boosting Machine (LGBM), Gradient Boosting (GB), Categorical Boosting (CB), Random Forest (RF), and Extra trees (ET). The construction of these models utilized the same features we had selected from our previous study whenever it was deemed best fit. During the model training process, we employed the k-fold cross validation.

K- fold and Cross Validation

Cross validation is often utilized when using ML models for the purpose of prediction as a way to identify how accurate a predictive model is performing in training.[34,35] When creating a ML model, an initial dataset labeled as the training dataset or training cohort is utilized as the source of data to train a ML model on carrying out its goal of prediction.[34] Once the model has been trained on this initial dataset or training cohort, the model is employed using a testing data set as a way of validation of its predictive ability.[34] As a way for evaluation, k-fold cross validation is utilized to process and identify the impact of a model's performance.[34] In this process a dataset is divided into K fold, where a fold is used once as the testing set for the data and the remaining folds are utilized as the training set of data.[34,36] This process is repeated until all datasets have been evaluated and the values for K-fold cross validation are represented with a mean score of the values within a ML model.[34] Based on prior research, a k value of 5 to 10 is believed to provide test error rate estimates without high biases or rates of variance.[34] For our data set we utilized the k-fold cross-validation and set our k from 7 to 10. Due to the imbalance of patients presenting for EDCA and non-EDCA patients, we utilized the Synthetic Minority Oversampling Technique (SMOTE) method.

Synthetic Minority Oversampling Technique (SMOTE)

In the setting of our data, evaluating the rates of EDCA based on large data plots of patients presenting to the ED, the value of the minority population (EDCA patients) is underrepresented in comparison to the majority (non-EDCA patients). Prior research has identified utility in the application of SMOTE. SMOTE works by creating synthetic values of the minority class from the provided training data within a ML model.[37] In essence, extra training data is created through operations on "real data" that allows for the creation of synthetic

examples that join each minority class data point.[37] The method of SMOTE allows for better classification and predictive performance of ML models as well as eliminating the need to under-sample the majority population of data points (non-EDCA patients).[37] For our research we applied the SMOTE technique with an oversampling of our sub-population of interest (EDCA patients) with an oversampling ratio of 0.6 times the majority population (non-EDCA patients). To evaluate the transparency and understanding of the connection of input to output of our constructed ML models, we utilized Shapley Additive exPlanations (SHAP).

Shapley Additive exPlanations (SHAP)

The importance of creating a model that can assess and make accurate predictions based on the data sets provided cannot be overstated. But as important as the creation and prediction of the model is the importance of being able to explain how the model can create its prediction based on each data set. SHAP provides for assignment of an importance value for a particular prediction based on the data sets entered.[38] From the data provided and the predictions made, SHAP technique allows for a unique solution to be made giving each data category considered within the training and testing cohort a value of importance to the solution at hand.[38] By utilizing SHAP in our data point, we are able to not only create our predictive models but also evaluate and interpret the importance of each feature selected for to construct our ML models and their predictions of patients at risk of EDCA events.[38] Lastly as a way to evaluate the performance and assess our metrics on the ML models ability to classify and predict accurately the occurrence of EDCA events, multitude of metrics were accounted for with definitions of interpretability listed.

Evaluation Metrics

- Specificity: The specificity of a test can best be defined in the setting of our ML models as the model's ability to predict accurately patients who are not at risk of having an EDCA event based on the data provided and analysis.[39]

- Negative predictive value (NPV): NPV in the setting of our test is the ability of our ML models to accurately predict patients who have the condition under study (EDCA) versus those who do not (non-EDCA patients). The ability of a predictive model to accurately screen patients and stratify their risk appropriately is incredibly important in the setting of

our ML models. A test with a strong NPV should be able to correctly predict a negative result in regards to risk stratification of those at risk of EDCA events.[39]

- Recall (sensitivity) The sensitivity of a test or predictive model is best defined as the ability of a test or predictive model to identify and detect correctly the true positive. In the case of our ML models, sensitivity is best reflected as the ability of our ML models to identify patients correctly who have an EDCA based on the data used to identify the condition.[39]

- Positive predictive value (PPV): Positive predictive value is strongly related to the sensitivity of a test. The positive predictive value of a test or ML model is best inferred as the percentage of patients predicted to have had an EDCA event who actually had an EDCA event based on the data provided.[39]

- F-1 score: In the field of machine learning, it is important to be able to evaluate and identify the ability of a ML models performance in the context of data classification.[40] As mentioned, in the development of ML models, a testing and training set is utilized to train and assess the performance of predictive power of a ML model.[40] The F-measure or F-1 score has been utilized to identify any issues in the ML models classification of the data used for making its predictions.[40] The F-1 score can be defined as the correctly classified data points expressed as a number as the correctly, incorrectly, or not classified objects being evaluated in a data set utilized by ML models.[40]

- Area under the receiver operative characteristic curve (AUC): The AUC is a common measure of the performance and predictive ability of ML models.[41] In terms of our ML models, the AUC is a predictive model that we use to identify whether a ML model is able to predict patients at risk of EDCA events based on the sensitivity and 1-specificity.[41] In terms of the area under the receiver operative characteristic curve, the AUC can be formed by drawing a straight line between each pairs of sensitivity and specificity which defines the AUC based on the sensitivity and 1-specificity.[41]

- Area under the precision recall curve (AUPRC): An AUPRC also known as a precision recall curve is closely similar to the AUC as an evaluator of the performance of a ML models predictive ability.[42] The utility of an AUPRC is that is can be utilized in ML model curves where there may be imbalances in the amount of observations or data points being studied (EDCA vs. non-EDCA patients). Prior research has identified AUPRC as a means to denote or identify a prediction or test and the results of patients with the presumed diagnosis (EDCA) versus those without the disease (non-EDCA) in a binary and understandable way.[42] The labeling of data points with use of AUPRC is those patients or data points with the disease of interest (EDCA) and those without the disease of interest; through this, a relationship in the ML models ability to predict accurately those with the disease process and the focus is on the true positive values within the data.[42] The relationship of an AUPRC is defined as the average of precision weighted by the probability of a given incident; for our research, this would be the ability of a test to make a prediction as compared to the probability of having an EDCA event.[42]

All our ML analyses took place with usage of Python 3.8 with the scikit-learn 0.23.1 package installed. Scikit-learn is a Python application designed for the utilization of supervised machine learning with emphasis on ease of use, performance, and API consistency.[43] The application was designed with the focus on allowing non-machine learning specialist to be able to perform tasks with the use of ML models with ease and efficiency.[43]

Results

A total of 237,349 patient cases were evaluated in the study based on the data from NTUH which served as our training set for our ML models. After screening for eligibility in effort to support our external validation based on the patient cases from BAS, 49,792 cases were included within our training data set. In total there were 477 (0.2%) and 166 (0.3%) cases which were identified as EDCA from the NTUH and BAS data sets, respectively. Our study population data sets and characteristics are represented within Table 1.

The characteristics and univariate analysis of features used to construct our ML models between patients with or without EDCA are listed within Table 2. In our prior study, a total of 54 clinical features were selected after setting the P value of less than 0.1 for the training cohort. These features included 14 triage data, 35 symptoms, as well as 5 demographics. In our study, we utilized a total of 41 features based on the prior models constructed which included 9 triage data, 30 laboratory results, as well as two demographics (age, sex).

Using K-fold cross-validation, with K ranges from 7 to 10, our constructed models demonstrated excellent ability to discriminate without the use of SMOTE both within the training and testing cohort. To be noted, the model built using the LGBM algorithm with a 7-fold approach achieved an AUC of 0.996 (95% Cl: 0.993-0.998) in the training cohort and an AUC of 0.897 (95% Cl: 0.876-0.916) in the testing cohort. No statistically significant differences were noted between the 6 ML models we had developed. Additional performance metrics, presenting the outcomes of the 6 distinct ML models without the incorporation of SMOTE are noted in Table 3 listed below.

With the application of the SMOTE technique the distinctions in AUC between the 6 ML models did not reach statistical significance. The LGBM algorithm was shown to have an AUC of 0.938 (95% CI: 0.929-0.948) with the training cohort and an AUC of 0.881 (95% CI:0.852-0.899) within the testing cohort. The outcomes from the 7-fold cross validation with features of performance for the 6 models are reflected in Table 1. Performance metrics for the classification

results with the application of SMOTE across K values of 7 to 10 of the 6 distinct ML models is listed in table 4 below.

The representation of each respective ML models functions as AUC both with and without the application of SMOTE is listed in Figure 1.

Figure 2 is a heat map of the features ordered by their median normalized importance in the predictive abilities of EDCA events both before the application of SMOTE (Figure 2A) and after SMOTE (Figure 2B). Of the features we utilized to train and test our ML models on, the top selected features included respiratory rate, consciousness, triage level, oxygen saturation, age, pulse rate, blood pressure, temperature, dyspnea, and pain severity.

Figure 3 is a visual construction using the SHAP values approach, where outcomes of each feature are analyzed, represented, and given a score for a specific prediction made using the constructed LGBM model. The model discussed was created through a 7-fold cross-validation process both with (Figure 3A) and without SMOTE (Figure 3B).

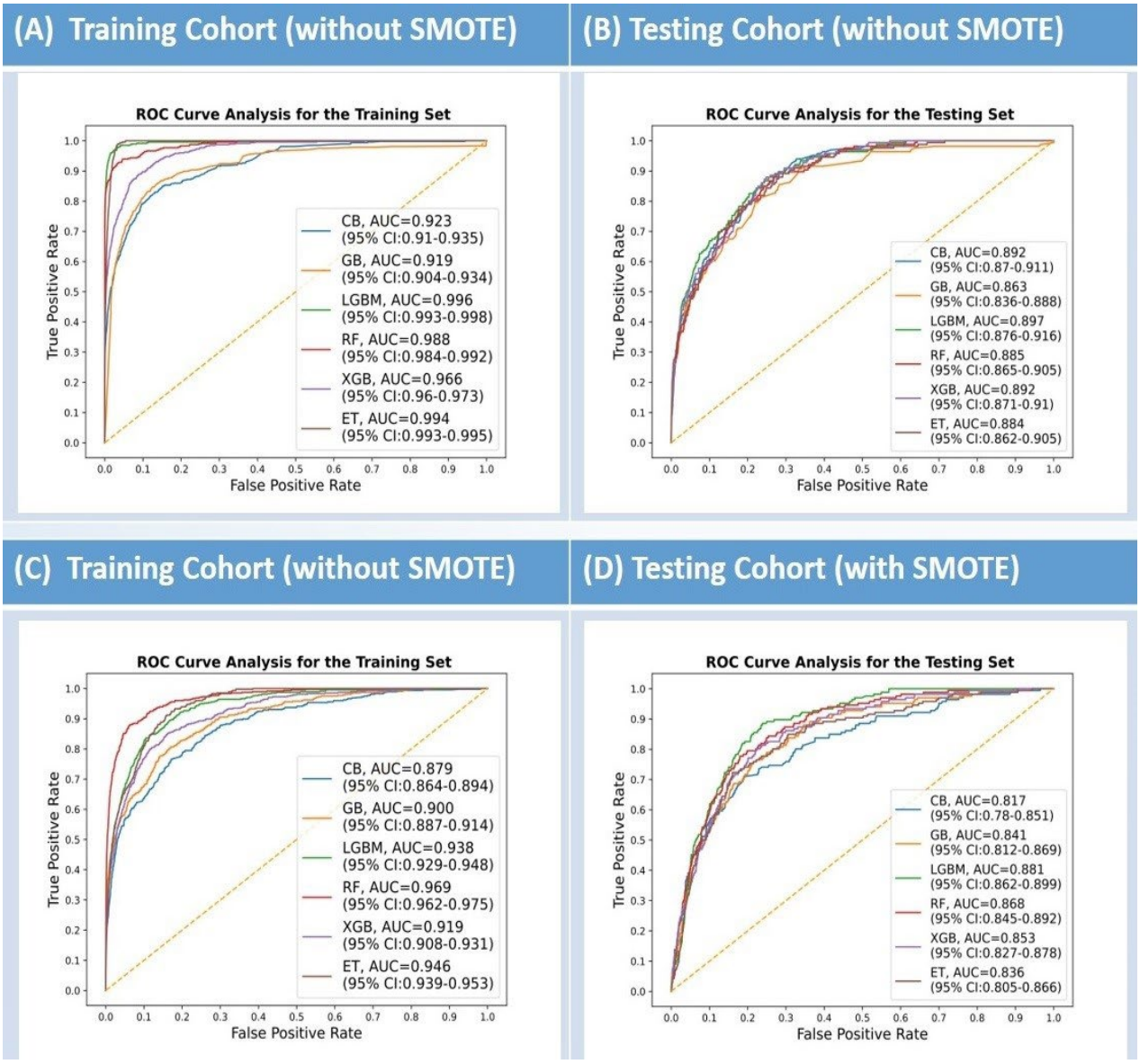

Figure 1. Comparative predictive powers of each of the six ML models in both the training (A), testing (B), training with SMOTE (C), and testing cohorts with SMOTE (D) based on their respective predictive abilities represented by the area under the receiver operating characteristics (AUPRC)

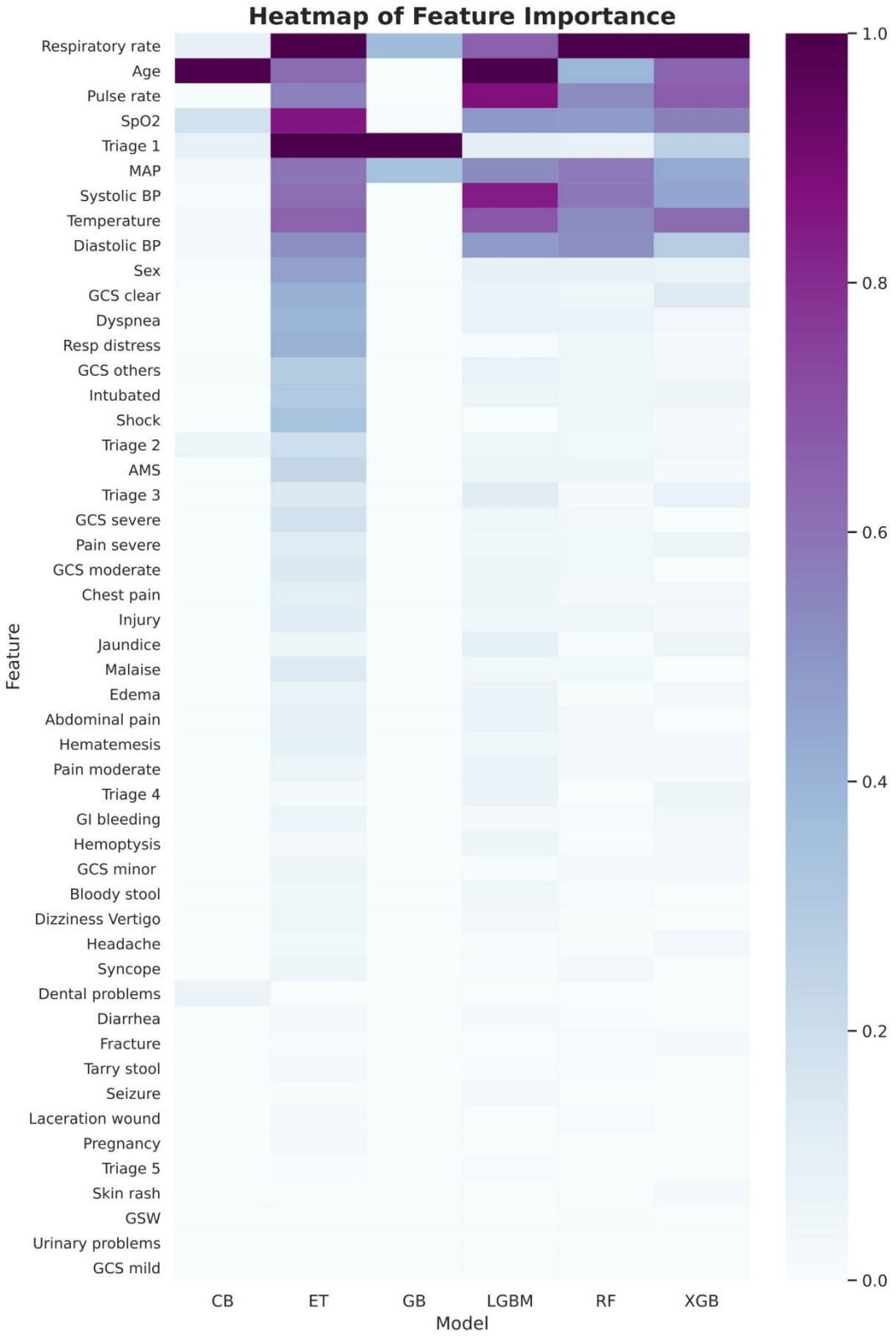

Figure 2A. Representation of feature importance for the 6 ML models prior to the application of SMOTE method.

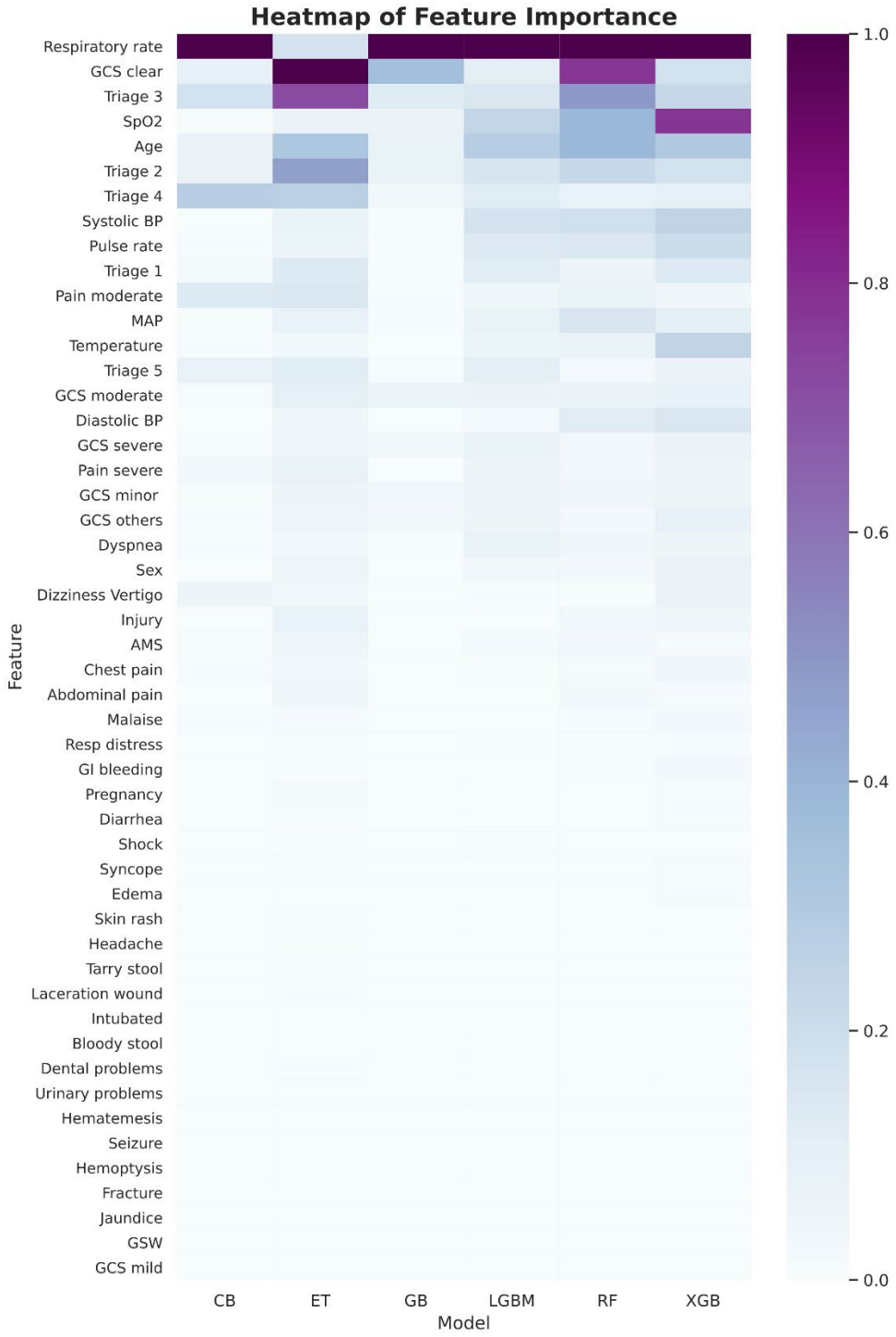

Figure 2B. Representation of feature importance for the 6 ML models after the application of SMOTE method.

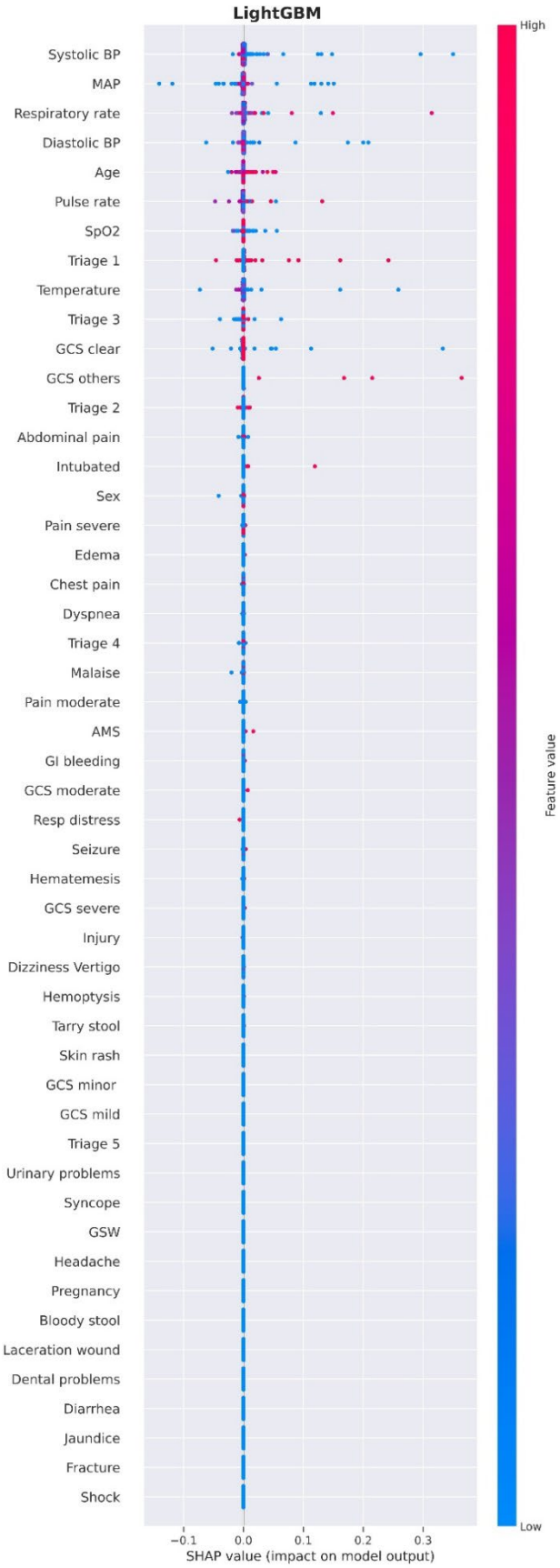

Figure 3A LGBM Without SMOTE

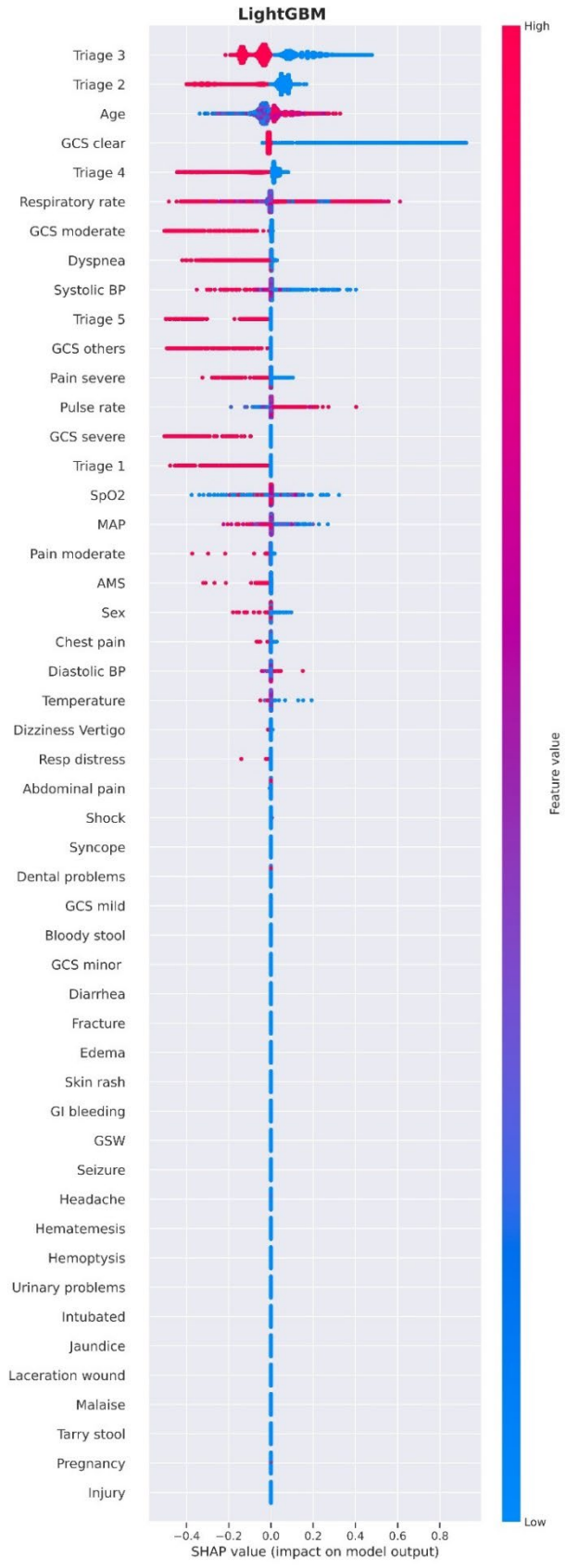

Figure 3B LGBM with SMOTE

Variables (Features)	Total (n= 287,141)	Training Cohort (n= 237,349)	Testing Cohort (n=49,792)	P value
Age, Mean (SD)	48.8 (19.9)	48.9 (20.0)	48.1 (19.6)	< 0.001
Systolic BP, Mean (SD)	135.9 (26)	136.2 (26.7)	134.7 (22.4)	< 0.001
Diastolic BP, Mean (SD)	80.0 (15.4)	80.5 (15.3)	77.8 (15.2)	<0.001
Mean Arterial Pressure, Mean (SD)	98.7 (17.2)	99.1 (17.5)	96.7 (15.5)	<0.001
Pulse Rate, Mean (SD)	88.7 (18.8)	89.1 (19.0)	86.7 (17.5)	<0.001
SpO2, Mean (SD)	97.4 (2.7)	97.1 (2.7)	98.4 (2.5)	<0.001
Respiratory Rate, Mean (SD)	18.1 (2.2)	18.2 (2.1)	17.8 (2.5)	<0.001
Temperature, Mean (SD)	98.5 (1.4)	98.5 (1.5)	98.5 (0.8)	<0.001
Sex				
Female	158111	126241	31870	
Male	129026	111108	17918	
Unknown	4	0	4	
GCS				
Clear	275376 (95.9)	227654 (95.9)	47722 (95.8)	
Mild	1464 (0.5)	0	1464 (2.9)	
Minor	1957 (0.7)	1957 (0.8)	0	
Moderate	4044 (1.4)	3712 (1.6)	332 (0.7)	
Others	2308 (0.8)	2145 (0.9)	163 (0.3)	
Severe	1992 (0.7)	1881 (0.8)	111 (0.2)	
AMS	5111 (1.8)	4163 (1.8)	948 (1.9)	0.02
Abdominal Pain	35595 (12.4)	29130 (12.3)	6465 (13.0)	<0.001
Bloody Stool	3874 (1.3)	3512 (1.5)	362 (0.7)	<0.001
Chest Pain	14893 (5.2)	11381 (4.8)	3512 (7.1)	<0.001
Dental Problems	4940 (1.7)	4564 (1.9)	376 (0.8)	<0.001
Diarrhea	7473 (2.6)	7084 (3.0)	389 (0.8)	<0.001
Dyspnea	17534 (6.1)	14252 (6.0)	3282 (6.6)	<0.001
Edema	1899 (0.7)	1076 (0.5)	823 (1.7)	<0.001
Fracture	4114 (1.4)	4106 (1.7)	8 (0.0)	<0.001
GI Bleed	4898 (1.7)	4426 (1.9)	472 (0.9)	<0.001
GSW	17 (0.0)	5 (0.0)	12 (0.0)	<0.001
Seizure	383 (0.1)	84 (0.0)	299 (0.6)	<0.001
Headache	10124 (3.5)	8883 (3.7)	1241 (2.5)	<0.001
Hematemesis	939 (0.3)	860 (0.4)	79 (0.2)	<0.001
Hemoptysis	870 (0.3)	846 (0.4)	24 (0.0)	<0.001
Injury	32994 (11.5)	32093 (13.5)	901 (1.8)	<0.001
Intubated	381 (0.1)	296 (0.1)	85 (0.2)	0.01
Jaundice	717 (0.2)	705 (0.3)	12 (0.0)	<0.001
Laceration Wound	8708 (3.0)	8402 (3.5)	306 (0.6)	<0.001
Malaise	4712 (1.6)	3729 (1.6)	983 (2.0)	<0.001
Pain Moderate	59822 (20.8)	58024 (24.4)	1798 (3.6)	<0.001
Pain Severe	36574 (12.7)	26663 (11.2)	9911 (19.9)	<0.001
Pregnancy	5256 (1.8)	4834 (2.0)	422 (0.8)	<0.001
Respiratory Distress	1519 (0.5)	1478 (0.6)	41 (0.1)	<0.001
Shock	1337 (0.5)	1180 (0.5)	157 (0.3)	<0.001
Skin Rash	4915 (1.7)	4566 (1.9)	349 (0.7)	<0.001
Syncope	1574 (0.5)	1455 (0.6)	119 (0.2)	<0.001
Urinary Problems	2986 (1.0)	2472 (1.0)	514 (1.0)	0.85
Dizziness/Vertigo	12524 (4.4)	11525 (4.9)	999 (2.0)	<0.001
Triage				
1	6768 (2.4)	6577 (2.8)	191 (0.4)	
2	70560 (24.6)	61011 (25.7)	9549 (19.2)	
3	174042 (60.6)	141637 (59.7)	32405 (65.1)	
4	29639 (10.3)	22415 (9.4)	7224 (14.5)	
5	6132 (2.1)	5709 (2.4)	423 (0.8)	

Table 1. Characteristics of the study population and the features utilized to train and test the ML models.

Training Cohort (n = 237,349)				Testing Cohort (n = 49,792)		
	EDCA (-) (n= 236,872)	EDCA (+) (n= 477)	P value	EDCA (-) (n= 49,626)	EDCA (+) (n=166)	P value
Age, Mean (SD)	48.9 (19.9)	68.6 (15.8)	<0.001	48.1 (19.5)	67.3 (13.5)	<0.001
Systolic BP, Mean (SD)	136.2 (26.7)	121.9 (34.5)	<0.001	134.7 (22.4)	125.3 (34.1)	<0.001
Diastolic BP, Mean (SD)	80.5 (15.3)	71.9 (19.8)	<0.001	77.8 (15.1)	73.8 (21.9)	<0.001
MAP, Mean (SD)	99.1 (17.5)	88.6 (23.4)	<0.001	96.8 (15.5)	91.0 (24.0)	<0.001
Pulse Rate, Mean (SD)	89.1 (19.0)	100.7 (28.2)	<0.001	86.7 (17.5)	93.2 (25.1)	<0.001
SpO2, Mean (SD)	97.1 (2.7)	93.3 (8.8)	<0.001	98.5 (2.4)	93.3 (11.1)	<0.001
Respiratory Rate, Mean (SD)	18.2 (2.1)	21.4 (4.6)	<0.0001	17.8 (2.5)	20.8 (5.5)	<0.001
Temperature, Mean (SD)	98.5 (1.4)	98.2 (2.4)	<0.001	98.5 (0.8)	98.2 (1.8)	<0.001
Sex			<0.001			<0.001
Female	126942 (53.2)	199 (41.7)		31797 (64.1)	73 (44.0)	
Male	110830 (46.8)	278 (58.3)		31797 (64.1)	73 (44.0)	
Unknown	0	0		4 (0.0)	0	
GCS			<0.001			<0.001
Clear	227343 (96.0)	311 (65.2)		47614 (95.9)	108 (65.1)	
Mild	0	0		1442 (2.9)	22 (13.3)	
Minor	1947 (0.8)	10 (2.1)		0	0	
Moderate	3665 (1.5)	47 (9.9)		330 (0.7)	2 (1.2)	
Others	2083 (0.9)	62 (13.0)		147 (0.3)	16 (9.6)	
Severe	1834 (0.8)	47 (9.9)		93 (0.2)	18 (10.8)	
AMS	4094 (1.7)	69 (14.5)	<0.001	927 (1.9)	21 (12.7)	<0.001
Abdominal Pain	29103 (12.3)	27 (5.7)	<0.001	6449 (13.0)	16 (9.6)	0.2
Bloody Stool	3498 (1.5)	14 (2.9)	0.008	362 (0.7)	0	0.27
Chest Pain	11350 (4.8)	31 (6.5)	0.08	3499 (7.1)	13 (7.8)	0.69
Dental Problems	4564 (1.9)	0	0.002	376 (0.8)	0	0.26
Diarrhea	7077 (3.0)	7 (1.5)	0.05	389 (0.8)	0	0.25
Dyspnea	14112 (6.0)	140 (29.4)	<0.001	3254 (6.6)	28 (16.9)	<0.001
Edema	1068 (0.5)	8 (1.7)	<0.001	823 (1.7)	0	0.09
Fracture	4104 (1.7)	2 (0.4)	0.03	8 (0.0)	0	0.87
GI Bleeding	4400 (1.9)	26 (5.5)	<0.001	471 (0.9)	1 (0.6)	0.65
GSW	4 (0.0)	1(0.2)	0.04	12 (0.0)	0	0.84
Seizure	83 (0.0)	1 (0.2)	0.04	299 (0.6)	0	0.32
Headache	8880 (3.7)	3 (0.6)	<0.001	1240 (2.5)	1 (0.6)	0.12
Hematemesis	847 (0.4)	13 (2.7)	<0.001	78 (0.2)	1 (0.6)	0.15
Hemoptysis	842 (0.4)	4 (0.8)	0.08	24 (0.0)	0	0.78
Injury	32072 (13.5)	21 (4.4)	<0.001	900 (1.8)	1 (0.6)	0.24
Intubated	275 (0.1)	21 (4.4)	<0.001	69 (0.1)	16 (9.6)	<0.001
Jaundice	698 (0.3)	7 (1.5)	<0.001	12 (0.0)	0	0.84
Laceration Wound	8399 (3.5)	3 (0.6)	<0.001	306 (0.6)	0	0.31
Malaise	3707 (1.6)	22 (4.6)	<0.001	971 (2.0)	12 (7.2)	<0.001
Pain moderate	58002 (24.5)	22 (4.6)	<0.001	1795 (3.6)	3 (1.8)	0.21
Pain severe	26626 (11.2)	37 (7.8)	0.02	9883 (19.9)	28 (16.9)	0.33
Pregnancy	4833 (2.0)	1 (0.2)	0.005	422 (0.9)	0	0.23
Respiratory Distress	1419 (0.6)	59 (12.4)	<0.001	33 (0.1)	8 (4.8)	<0.001
Shock	1141 (0.5)	39 (8.2)	<0.001	152 (0.3)	5 (3.0)	<0.001
Skin Rash	4566 (1.9)	0	0.002	349 (0.7)	0	0.28
Syncope	1448 (0.6)	7 (1.5)	0.02	119 (0.2)	0	0.53
Tarry Stool	3660 (1.5)	15 (3.1)	0.005	24 (0.0)	0	0.78
Urinary Problems	2472 (1.0)	0 (0.0)	0.02	513 (1.0)	1 (0.6)	0.58
Dizziness/Vertigo	11513 (4.9)	12 (2.5)	0.02	997 (2.0)	2 (1.2)	0.46
Triage			<0.001			<0.001
1	6388 (2.7)	189 (39.6)		160 (0.3)	31 (18.7)	
2	60825 (25.7)	186 (39.0)		9456 (19.1)	93 (56.0)	
3	141540 (59.8)	97 (20.3)		32363 (65.2)	42 (25.3)	
4	22411 (9.5)	4 (0.8)		7224 (14.6)	0 (0.0)	
5	5708 (2.4)	1 (0.2)		423 (0.9)	0 (0.0)	

Table 2. Characteristics and univariate analysis of features selected between patients with EDCA and without EDCA events in the training and testing cohorts respectively.

Classifier	Cohort	k fold	AUC (95% CI)	AUPRC (95% CI)	Accuracy	F1	Kappa	Sensitivity	Specificity	PPV	NPV
RF	test	7	0.885 (0.865-0.905)	0.086 (0.052-0.13	0.997	0	0	0	1	NaN	0.997
GB	test	7	0.863 (0.836-0.888)	0.065 (0.042-0.09	0.995	0.181	0.179	0.169	0.998	0.196	0.997
ET	test	7	0.884 (0.862-0.905)	0.091 (0.057-0.14	0.997	0	0	0	1	NaN	0.997
XGB	test	7	0.892 (0.871-0.91)	0.081 (0.049-0.12	0.997	0.024	0.024	0.012	1	1	0.997
LGBM	test	7	0.897 (0.876-0.916)	0.082 (0.05-0.123	0.997	0.058	0.058	0.03	1	0.833	0.997
CB	test	7	0.892 (0.87-0.911)	0.077 (0.05-0.119	0.997	0.012	0.012	0.006	1	1	0.997
RF	test	8	0.885 (0.865-0.905)	0.089 (0.053-0.13	0.997	0	0	0	1	NaN	0.997
GB	test	8	0.857 (0.828-0.882)	0.065 (0.042-0.09	0.995	0.181	0.178	0.169	0.998	0.194	0.997
ET	test	8	0.884 (0.862-0.905)	0.091 (0.057-0.14	0.997	0	0	0	1	NaN	0.997
XGB	test	8	0.892 (0.871-0.91)	0.081 (0.049-0.12	0.997	0.024	0.024	0.012	1	1	0.997
LGBM	test	8	0.897 (0.876-0.916)	0.082 (0.05-0.123	0.997	0.058	0.058	0.03	1	0.833	0.997
CB	test	8	0.887 (0.865-0.907)	0.083 (0.049-0.12	0.997	0.024	0.024	0.012	1	1	0.997
RF	test	9	0.884 (0.862-0.905)	0.085 (0.051-0.13	0.997	0	0	0	1	NaN	0.997
GB	test	9	0.857 (0.828-0.882)	0.065 (0.042-0.09	0.995	0.181	0.178	0.169	0.998	0.194	0.997
ET	test	9	0.876 (0.851-0.899)	0.097 (0.061-0.14	0.997	0	0	0	1	NaN	0.997
XGB	test	9	0.896 (0.876-0.915)	0.086 (0.054-0.13	0.997	0	0	0	1	NaN	0.997
LGBM	test	9	0.897 (0.876-0.916)	0.082 (0.05-0.123	0.997	0.058	0.058	0.03	1	0.833	0.997
CB	test	9	0.887 (0.866-0.908)	0.078 (0.046-0.12	0.997	0.012	0.012	0.006	1	1	0.997
RF	test	10	0.885 (0.865-0.905)	0.089 (0.053-0.13	0.997	0	0	0	1	NaN	0.997
GB	test	10	0.862 (0.836-0.886)	0.065 (0.042-0.09	0.995	0.18	0.178	0.169	0.998	0.193	0.997
ET	test	10	0.876 (0.852-0.9)	0.097 (0.061-0.14	0.997	0	0	0	1	NaN	0.997
XGB	test	10	0.892 (0.871-0.91)	0.081 (0.049-0.12	0.997	0.024	0.024	0.012	1	1	0.997
LGBM	test	10	0.897 (0.876-0.916)	0.082 (0.05-0.123	0.997	0.058	0.058	0.03	1	0.833	0.997
CB	test	10	0.892 (0.87-0.911)	0.077 (0.05-0.119	0.997	0.012	0.012	0.006	1	1	0.997
RF	train	7	0.988 (0.984-0.992)	0.805 (0.773-0.83	0.998	0.312	0.311	0.184	1	1	0.998
GB	train	7	0.919 (0.904-0.934)	0.038 (0.032-0.04	0.983	0.095	0.092	0.44	0.984	0.053	0.999
ET	train	7	0.994 (0.993-0.995)	0.348 (0.308-0.39	0.998	0	0	0	1	NaN	0.998
XGB	train	7	0.966 (0.96-0.973)	0.439 (0.396-0.48	0.998	0.391	0.39	0.243	1	0.991	0.998
LGBM	train	7	0.996 (0.993-0.998)	0.732 (0.692-0.76	0.998	0.345	0.345	0.212	1	0.935	0.998
CB	train	7	0.923 (0.91-0.935)	0.324 (0.282-0.36	0.998	0.375	0.374	0.231	1	1	0.998
RF	train	8	0.987 (0.983-0.991)	0.793 (0.761-0.82	0.998	0.312	0.311	0.184	1	1	0.998
GB	train	8	0.922 (0.906-0.936)	0.039 (0.033-0.04	0.983	0.094	0.091	0.438	0.984	0.053	0.999
ET	train	8	0.994 (0.993-0.995)	0.348 (0.308-0.39	0.998	0	0	0	1	NaN	0.998
XGB	train	8	0.966 (0.96-0.973)	0.439 (0.396-0.48	0.998	0.391	0.39	0.243	1	0.991	0.998
LGBM	train	8	0.996 (0.993-0.998)	0.732 (0.692-0.76	0.998	0.345	0.345	0.212	1	0.935	0.998
CB	train	8	0.925 (0.913-0.937)	0.346 (0.301-0.38	0.998	0.378	0.377	0.233	1	1	0.998
RF	train	9	0.972 (0.963-0.979)	0.669 (0.628-0.70	0.998	0.133	0.133	0.071	1	1	0.998
GB	train	9	0.922 (0.906-0.936)	0.039 (0.033-0.04	0.983	0.094	0.091	0.438	0.984	0.053	0.999
ET	train	9	0.992 (0.991-0.993)	0.279 (0.243-0.32	0.998	0	0	0	1	NaN	0.998
XGB	train	9	0.959 (0.95-0.966)	0.396 (0.353-0.43	0.998	0.375	0.374	0.231	1	1	0.998
LGBM	train	9	0.996 (0.993-0.998)	0.732 (0.692-0.76	0.998	0.345	0.345	0.212	1	0.935	0.998
CB	train	9	0.924 (0.912-0.937)	0.341 (0.297-0.38	0.998	0.378	0.377	0.233	1	1	0.998
RF	train	10	0.987 (0.983-0.991)	0.793 (0.761-0.82	0.998	0.312	0.311	0.184	1	1	0.998
GB	train	10	0.92 (0.904-0.935)	0.038 (0.033-0.04	0.983	0.094	0.091	0.438	0.984	0.053	0.999
ET	train	10	0.992 (0.991-0.993)	0.283 (0.246-0.32	0.998	0	0	0	1	NaN	0.998
XGB	train	10	0.966 (0.96-0.973)	0.439 (0.396-0.48	0.998	0.391	0.39	0.243	1	0.991	0.998
LGBM	train	10	0.996 (0.993-0.998)	0.732 (0.692-0.76	0.998	0.345	0.345	0.212	1	0.935	0.998
CB	train	10	0.923 (0.91-0.935)	0.324 (0.282-0.36	0.998	0.375	0.374	0.231	1	1	0.998

Table 3. Data and analysis of ML models with a k-fold cross-validation between 7 and 10 without SMOTE.

Classifier	Cohort	k fold	AUC (95% CI)	AUPRC (95% CI)	Accuracy	F1	Kappa	Sensitivity	Specificity	PPV	NPV
RF	test	7	0.868 (0.845-0.892)	0.025 (0.018-0.04	0.987	0.052	0.047	0.108	0.99	0.034	0.997
GB	test	7	0.841 (0.812-0.869)	0.023 (0.017-0.03	0.988	0.054	0.05	0.102	0.991	0.037	0.997
ET	test	7	0.836 (0.805-0.866)	0.017 (0.013-0.02	0.979	0.029	0.024	0.096	0.982	0.017	0.997
XGB	test	7	0.853 (0.827-0.878)	0.023 (0.017-0.03	0.982	0.058	0.052	0.169	0.984	0.035	0.997
LGBM	test	7	0.881 (0.862-0.899)	0.021 (0.016-0.02	0.967	0.041	0.035	0.211	0.97	0.023	0.997
CB	test	7	0.817 (0.78-0.851)	0.027 (0.018-0.04	0.976	0.058	0.052	0.223	0.978	0.033	0.997
RF	test	8	0.87 (0.846-0.893)	0.026 (0.018-0.04	0.987	0.055	0.05	0.114	0.99	0.036	0.997
GB	test	8	0.841 (0.812-0.869)	0.023 (0.017-0.03	0.988	0.054	0.05	0.102	0.991	0.037	0.997
ET	test	8	0.836 (0.806-0.866)	0.017 (0.013-0.02	0.979	0.029	0.024	0.096	0.982	0.017	0.997
XGB	test	8	0.853 (0.827-0.878)	0.023 (0.017-0.03	0.982	0.058	0.052	0.169	0.984	0.035	0.997
LGBM	test	8	0.881 (0.862-0.899)	0.021 (0.016-0.02	0.967	0.041	0.035	0.211	0.97	0.023	0.997
CB	test	8	0.817 (0.78-0.851)	0.027 (0.018-0.04	0.976	0.058	0.052	0.223	0.978	0.033	0.997
RF	test	9	0.87 (0.846-0.893)	0.026 (0.018-0.04	0.987	0.055	0.05	0.114	0.99	0.036	0.997
GB	test	9	0.841 (0.812-0.869)	0.023 (0.017-0.03	0.988	0.054	0.05	0.102	0.991	0.037	0.997
ET	test	9	0.836 (0.805-0.866)	0.017 (0.013-0.02	0.979	0.029	0.024	0.096	0.982	0.017	0.997
XGB	test	9	0.853 (0.827-0.878)	0.023 (0.017-0.03	0.982	0.058	0.052	0.169	0.984	0.035	0.997
LGBM	test	9	0.881 (0.862-0.899)	0.021 (0.016-0.02	0.967	0.041	0.035	0.211	0.97	0.023	0.997
CB	test	9	0.817 (0.78-0.851)	0.027 (0.018-0.04	0.976	0.058	0.052	0.223	0.978	0.033	0.997
RF	test	10	0.87 (0.846-0.893)	0.026 (0.018-0.04	0.987	0.055	0.05	0.114	0.99	0.036	0.997
GB	test	10	0.841 (0.812-0.869)	0.023 (0.017-0.03	0.988	0.054	0.05	0.102	0.991	0.037	0.997
ET	test	10	0.836 (0.805-0.866)	0.017 (0.013-0.02	0.979	0.029	0.024	0.096	0.982	0.017	0.997
XGB	test	10	0.853 (0.827-0.878)	0.023 (0.017-0.03	0.982	0.058	0.052	0.169	0.984	0.035	0.997
LGBM	test	10	0.881 (0.862-0.899)	0.021 (0.016-0.02	0.967	0.041	0.035	0.211	0.97	0.023	0.997
CB	test	10	0.817 (0.78-0.851)	0.027 (0.018-0.04	0.976	0.058	0.052	0.223	0.978	0.033	0.997
RF	train	7	0.969 (0.962-0.975)	0.305 (0.259-0.35	0.997	0.33	0.329	0.405	0.998	0.279	0.999
GB	train	7	0.9 (0.887-0.914)	0.231 (0.188-0.27	0.998	0.294	0.293	0.224	0.999	0.425	0.998
ET	train	7	0.946 (0.939-0.953)	0.104 (0.078-0.13	0.996	0.158	0.156	0.189	0.998	0.136	0.998
XGB	train	7	0.919 (0.908-0.931)	0.128 (0.097-0.16	0.996	0.193	0.192	0.226	0.998	0.169	0.998
LGBM	train	7	0.938 (0.929-0.948)	0.134 (0.103-0.16	0.994	0.153	0.151	0.268	0.995	0.107	0.999
CB	train	7	0.879 (0.864-0.894)	0.081 (0.059-0.11	0.992	0.112	0.11	0.249	0.994	0.073	0.998
RF	train	8	0.969 (0.962-0.975)	0.307 (0.262-0.35	0.997	0.33	0.329	0.405	0.998	0.279	0.999
GB	train	8	0.9 (0.887-0.914)	0.231 (0.188-0.27	0.998	0.294	0.293	0.224	0.999	0.425	0.998
ET	train	8	0.946 (0.939-0.953)	0.104 (0.078-0.13	0.996	0.16	0.158	0.189	0.998	0.138	0.998
XGB	train	8	0.919 (0.908-0.931)	0.128 (0.097-0.16	0.996	0.193	0.192	0.226	0.998	0.169	0.998
LGBM	train	8	0.938 (0.929-0.948)	0.134 (0.103-0.16	0.994	0.153	0.151	0.268	0.995	0.107	0.999
CB	train	8	0.879 (0.864-0.894)	0.081 (0.059-0.11	0.992	0.112	0.11	0.249	0.994	0.073	0.998
RF	train	9	0.969 (0.962-0.975)	0.307 (0.262-0.35	0.997	0.33	0.329	0.405	0.998	0.279	0.999
GB	train	9	0.9 (0.887-0.914)	0.231 (0.188-0.27	0.998	0.294	0.293	0.224	0.999	0.425	0.998
ET	train	9	0.946 (0.939-0.953)	0.104 (0.078-0.13	0.996	0.158	0.156	0.189	0.998	0.136	0.998
XGB	train	9	0.919 (0.908-0.931)	0.128 (0.097-0.16	0.996	0.193	0.192	0.226	0.998	0.169	0.998
LGBM	train	9	0.938 (0.929-0.948)	0.134 (0.103-0.16	0.994	0.153	0.151	0.268	0.995	0.107	0.999
CB	train	9	0.879 (0.864-0.894)	0.081 (0.059-0.11	0.992	0.112	0.11	0.249	0.994	0.073	0.998
RF	train	10	0.969 (0.962-0.975)	0.307 (0.262-0.35	0.997	0.33	0.329	0.405	0.998	0.279	0.999
GB	train	10	0.9 (0.887-0.914)	0.231 (0.188-0.27	0.998	0.294	0.293	0.224	0.999	0.425	0.998
ET	train	10	0.946 (0.939-0.953)	0.104 (0.078-0.13	0.996	0.158	0.156	0.189	0.998	0.136	0.998
XGB	train	10	0.919 (0.908-0.931)	0.128 (0.097-0.16	0.996	0.193	0.192	0.226	0.998	0.169	0.998
LGBM	train	10	0.938 (0.929-0.948)	0.134 (0.103-0.16	0.994	0.153	0.151	0.268	0.995	0.107	0.999
CB	train	10	0.879 (0.864-0.894)	0.081 (0.059-0.11	0.992	0.112	0.11	0.249	0.994	0.073	0.998

Table 4. Data and analysis of ML models with a k-fold cross-validation between 7 and 10 with SMOTE.

Discussion

Through our cross-country approach for external validation, our ML models showed strong predictive performance as recognized by the AUC, specificity, as well as NPV. Values for AUC at a k-fold cross validation of 7 within the testing cohort showed an AUC for each model of 0.868 (RF), 0.841 (GB), 0.836 (ET), 0.853 (XGB), 0.881 (LGBM), and 0.817 (XGB). In addition, the NPV for all 6 of our ML models at a k-fold cross-validation of 7 was 0.997; this value shows strong ability of our ML models to discriminate patients without EDCA events based on the selected features. Although as previously mentioned, there was not a statistically significant difference between the individual ML models discriminatory abilities, all models performed with excellent discriminatory abilities to predict patients at risk of EDCA events based on the data presented. Our models were not only able to discriminate against patients who were not at risk of EDCA events within the training and testing cohort but were also able to predict patients at risk of EDCA events based on the AUC listed above.

The focus of our study was on the validation of our ML models in a diverse population outside of the initial testing population of NTUH ED. The importance of showing external validation in a cross-country setting provides for stronger generalization of ML models and their ability to predict EDCA across a diverse range of populations. Through the results of our study on ML models' ability to provide predictive evidence of patients at risk of EDCA, we hope to provide evidence to support their implementation in the EMR system as a healthcare tool in the ED for resource allocation, triage, management, as well as risk stratification for patients at risk of IHCA events in the ED.

Comparison From Initial Study

Our aim of this cross-country study was to build off our initial study and further provide evidence of validity to the utility of our ML Models ability to predict EDCA. In our initial study we compared the predictive ability of our ML models to that of the NEWS 2 score. In the initial study, 3 ML models were constructed (ET, GBM, and RF).[26] Similarly, the previous study also showed non-statistically significant differences between the 3 constructed models but all models showed excellent performance in prediction of EDCA based on the AUC (ET 0.915, GBM, 0.930, and RF of 0.931).[26] In addition, our initial study ML models also significantly outperformed the NEWS 2 scoring system based on the AUC (AUC of 0.678).[26] In addition, we

built off of our initial parameters for ML models by re-defining the features selected for by setting the P value to less than 0.1. By doing so, we accounted for a total of 41 features versus the initial 54 clinical features. With the application of SMOTE to our data, we identified that the topmost important features in prediction of EDCA events were respiratory rate, consciousness, triage level, oxygen saturation, age, pulse rate, blood pressure, temperature, dyspnea, and pain severity. This highlights one of the most interesting aspects of the utility of supervised learning with ML models, the ability to draw associations and levels of importance to input data on the output or outcome of interest. In the setting of acute events such as cardiac arrest the importance of a strong predictive model is not only on its effectiveness of prediction but also on understanding how it is able to make the predictions that it does. In previous studies of ML models and artificial intelligence, a common limitation brought up is lack of understanding of how ML models can make their prediction. To provide strong understanding of the ML models rationale, we employed the SHAP method; To avoid the need to under-sample the majority population of each cohort (non-EDCA patients) we utilized the SMOTE method.

Model Interpretation and Explainability

- Shapley Additive explanation (SHAP)

When constructing a ML model, the importance of not only the outcome but also on how each feature correlates to the predictions made is of the utmost importance and cannot be overstated. Not only is it necessary for a ML model to utilize inputs to draw a connection to the outputs of interest (EDCA events), but also how each feature plays a part in the conclusion drawn. For that reason, we utilized the SHAP method. The SHAP is a formula which can provide a formula to assign each feature a value of importance or contribution to the sum or outcome of interest.[44] Prior research has identified that SHAP uses an additive contribution method which creates a linear model of the importance of feature.[45] By utilizing SHAP we gain understanding not only in the interpretation of the prediction but the significance of each individual feature in making the predictions of EDCA. The utility of SHAP for our data on showing the importance of each feature selection provides for further evidence and opportunity for refinement of our current ML models to increase their predictive power. As previously mentioned, we refined our initial studies 54 clinical features

through a P value of 0.1 to 41 features in our external validation study and created a heat map (Figure 3A and 3B) to give a representation of the importance of each feature. In addition to the utility SHAP offered our data to assign feature importance in a quantitative approach, we employed the SMOTE method to account for the limitations of our studies EDCA population.

- Synthetic Minority Oversampling Technique (SMOTE)

  SMOTE is a statistical analysis tool that allows for the evaluation of non-uniform data. The method of SMOTE allows for the creation of synthetic data for the minority population (EDCA patients) without the need for under-sampling of the majority population (non-EDCA patients). In the setting of our data it allows for better classification and predictive performance.[37] We utilized a SMOTE with an oversampling ratio of 0.6 times the majority population; at this rate our assumption is that the EDCA population would be as representative of the true normal in the general population for the purposes of external validation. At this rate there was an increase in the sensitivity of each of the ML models noted with SMOTE in the predictive ability of each respective ML model. For example, LGBM sensitivity both without and with SMOTE was 0.03 and 0.211 in the testing cohort with k-fold cross-validation of 7.

Limitations

A limitation of our study can be identified in the lack of satisfactory sensitivity and positive predictive value (PPV). We attribute these findings to the question and study at hand of creation and external validation of ML models for the prediction of EDCA events. Additionally, it was noted that with the use of SMOTE there was a noticeable improvement in the model's sensitivity and predictive power. A follow up to this limitation may be further studies with larger cross-country ED populations to continue to prove utility of the ML models as a predictive tool within the ED.

Another concern or limitation of our study may be the variability of data and concern for biases in the multi-country dataset. To address this in the future, follow-up study using data sets from multiple EDs across the country for normalization may be appropriate. In addition, sensitivity (recall) of the models was noted to have varied significantly between our initial study and the current cross-country study. In our initial study, sensitivity at a k-fold cross-validation of

9 for RF, GB, and ET was noted to be 0.748, 0.736, and 0.761 within the testing cohort respectively.[26] In our current study at the same cross-validation value the sensitivity for ML models RF, GB, and ET were noted to be 0.114, 0.102, and 0.096 respectively. The concern for causes of the decline in sensitivity may be to the differences within the populations sampled as well as the change in features selected for training and testing of our 6 ML models in the external validation population. To address these changes, follow up study with additional demographics, triage data, as well as laboratory values may be appropriate for increasing the predictive power of our ML models. In addition, through the application of SHAP at that stage, more refinement can be made for follow up studies in terms of feature selection as deemed best fit.

Clinical Application

The goal of our cross-country external validation study was to identify the utility, predictive power, and effectiveness of our ML models as a clinical tool to be implemented in EDs globally as a predictive tool to offer advancement in patient care and risk stratification for patients at risk of EDCA. Through our previous study, we identified that in comparison to NEWS 2 score, our ML models were much more effective in predicting the risk of EDCA. We believe by offering a stronger predictive tool to EDs for patients at risk of EDCA we can support healthcare workers in multiple different ways.

Patients at risk of EDCA are vulnerable due to long boarding times, the variance of recognition and observation of vital signs as well as high patient volumes.[46–50] Without the appropriate recognition and management, patients who show signs of vital sign instability remain at risk of EDCA.[46] Previous research has discussed the importance of earlier identification of patients at risk of EDCA events for the rapid facilitation of appropriate interventions.[46]

We hope that our ML models and their predictive performance shows strength as a useful tool for implementation within emergency departments globally to increase efficiency, allocation of resources, and prevent adverse events from occurring through early identification of vital sign instability. Through implementation in the clinical setting, our ML models can not only provide for early identification, but also learn simultaneously from the constant data how to better predict EDCA events.

Clinical limitations for our ML models and their implementation include a possible learning curve for health care personnel. The most important aspect of incorporating a new

system within a hospital system would be the training of health care personnel on how to best utilize it. The initial learning would be on understanding how to identify warning scores or patient EDCA risk stratification based on the ML models. The following learning would be on designating resource allocation to patients who the ML models identify at greatest risk of having EDCA events. The goal would be for all patients who present to the ED to have the ML algorithms processing the inputted data from triage in conjunction with demographics, laboratory results, and chief concerns while they are undergoing their initial work up in the ED. In addition, there may be concerns of over-sensitivity which could lead to fatigue from alerts being prompted in the ED from the ML models. One way we feel this may be mitigated again reflects on the design of supervised learning.

Through constant data entry and processing, the ML models can learn and become more able to stratify patients at risk of EDCA events. The ability of artificial intelligence to grow from larger data processing provides ample utility in their predictive modeling learning and usefulness as a clinical tool. This leads into the discussion of the innovation of not only our own research on ML algorithms in prediction of EDCA, but the use of ML as a clinical tool for risk stratification across different disease processes in the healthcare setting.

Innovation

The innovation of our research not only stems from its ability to provide predictive risk stratification to a subset patient population within the ED at risk of IHCA events (EDCA). Our initial study comparing 3 ML models to the predictive power of the NEWS 2 score showed strong discriminatory strength in assessing patients at risk and not at risk of EDCA events based on statistically significant AUC's in comparison to one of the current EWS systems (NEWS 2). Our goal of our current research was to build off our previous findings through study of the ML models in a diverse population outside of that which we utilized in our initial study (NTUH).

Through our findings we have established that our ML models continue to show strong discriminatory strength in risk stratification and prediction of EDCA events in a cross-country approach using a different population from our initial study. In addition, the utility of ML models as an application in healthcare is highlighted as a modality for possible incorporation into the EMR system of hospitals on a global setting. In the past few years, machine learning and

artificial intelligence has gained significant interest in both the healthcare industry as well as others. Through its ability of supervised learning, these models are not only able to provide prediction, but also draw conclusions from inputs to results through their processing of large data volumes. As new studies continue to explore the utility of ML models in the hospital setting, we hope to add to the current data and provide evidence of their possible future within the EMR systems of hospitals on a global scale.

Future Directions

Our external validation study has provided evidence in support of the ML models and their ability to predict patients at risk of EDCA events based on triage data and clinical features selected for. Our hope for future studies lies in how to further improve their ability to identify patients at risk of EDCA. Possible future studies would include incorporation of more data features to the training and testing cohort such as laboratory data, triage data, and other vital signs with hope of increasing their sensitivity to EDCA events. In addition, our current study for external validation took place at a tertiary teaching hospital within North Texas. We hope through utilization of a possible multi-center approach with much more significant ED data, we would provide the best fit population for testing of our ML models to provide ample evidence of their predictive properties as an ED tool for EDCA evaluation.

Other areas for future research would be the real-time usage of our ML models in an ED to evaluate their clinical amplitude and response times to EDCA events. By providing and testing our model through a longitudinal study, we can focus on parameters such as hospital stay, mortality, resource utilization, STEMI alerts, and rapid response times both pre-ML model implementation and post-implementation. This would build off the validation study we have provided and give evidence of its real-time effect on the efficiency and effectiveness as a triage tool in the hospital setting.

Through evidence of our ML models predictive abilities in the setting the EDCA, future directions of study may be on other disease processes and avenues of enhancing patient care through use of supervised learning algorithms. Beginning with EDCA, future avenues for ML models and their utility could be in the prediction of other disease processes with current warning systems in place such as ICU admission risk stratification, sepsis risk stratification, or

cerebrovascular event stratification. The field of ML with the use of large data amounts to draw conclusions from initial inputs based on patient features provides for room for enhancing the current field of medicine through supervised learning in real-time of patients who require higher level of care. We hope that our initial project provides evidence of its utility and room for improvement and further innovation in its application in the healthcare setting.

Conclusions

The purpose of our research study was to build off our previous study validating the use of ML models trained and tested within a hospital setting at the National Taiwan University Hospital to predict patients at risk of EDCA. We sought to train and test our ML models utilizing triage features and clinical data provided from an emergency department within the US at Baylor All Saints in Fort Worth Texas. After data analysis, our 6 ML models showed excellent predictive performance in AUC, specificity, and NPV as can be identified in the results section above. By providing for external validation, we offer more reliability of our ML models to be adopted and utilized as a clinical tool in the prediction and risk stratification of patients who present to the ED with concerns of possible EDCA events. The cross-country approach provided an ample opportunity to demonstrate the ML models predictive ability in a diverse population outside of our initial population in Taiwan.

Through our external validation study, we hope to provide an example of how predictive modeling with use of supervised ML models can not only improve patient outcomes but workflow within the emergency department. With implementation of ML models for their predictive abilities, we can offer an efficient and reliable tool to improve resource utilization, triage mechanisms, and alert systems for health care providers to identify and respond to most effectively.

Using the SHAP method, we are also able to provide evidence of the significance of each feature selected for to construct our ML models and their predictions of EDCA events. Our hope for future research would be to increase the features used to train and test our cohorts to further improve their predictive ability of EDCA events. Such features to include for future research may include imaging findings and impressions, laboratory data values, as well as EKG impressions, and other demographics.

Our final goal for future research based on the external validation study we have conducted would be to implement the use of ML models in a healthcare setting and provide longitudinal evidence of its utility and rationale in predictive modeling of EDCA events. Our hope would be that through evidence of its utility in the ED to predict at-risk patients, we can create more efficient hospital protocols in place to respond accordingly with adequate personnel and resources to provide the best care for patients at risk of cardiac arrest.

Compliance

All adult patients who presented to the BAS ED during the time listed were identified in the study. The study was approved by the Institutional Review Board of NTUH (201606072RINA) and BAS (reference number: 344143) and waived the requirements for informed consent.

References

1. Andersen LW, Holmberg MJ, Berg KM, Donnino MW, Granfeldt A. In-Hospital Cardiac Arrest: A Review. *JAMA*. 2019;321(12):1200-1210. doi:10.1001/jama.2019.1696
2. Penketh J, Nolan JP. In-hospital cardiac arrest: the state of the art. *Crit Care*. 2022;26(1):376. doi:10.1186/s13054-022-04247-y
3. Andersson A, Arctaedius I, Cronberg T, et al. In-hospital versus out-of-hospital cardiac arrest: Characteristics and outcomes in patients admitted to intensive care after return of spontaneous circulation. *Resuscitation*. 2022;176:1-8.
4. Penketh J, Nolan JP. In-hospital cardiac arrest: the state of the art. *Crit Care*. 2022;26(1):376. doi:10.1186/s13054-022-04247-y
5. Nolan JP, Berg RA, Andersen L. Utstein Collaborators: Cardiac arrest and cardiopulmonary resuscitation outcome reports: Update of the utstein resuscitation registry template for in-hospital cardiac arrest: A consensus report from a task force of the international liaison committee on resuscitation (American Heart Association, European Resuscitation Council, Australian and New Zealand Council on Resuscitation, Heart and Stroke Foundation of canada, Interamerican Heart Foundation, Resuscitation Council of Southern Africa, Resuscitation Council of Asia). *Resuscitation*. 2019;144:166-177.
6. Nolan JP, Sandroni C, Böttiger BW, et al. European resuscitation council and European society of intensive care medicine guidelines 2021: post-resuscitation care. *Resuscitation*. 2021;161:220-269.
7. Soar J, Böttiger BW, Carli P, et al. European resuscitation council guidelines 2021: adult advanced life support. *Resuscitation*. 2021;161:115-151.
8. Moskowitz A, Holmberg MJ, Donnino MW, Berg KM. In-Hospital Cardiac Arrest–Are We Overlooking a Key Distinction? *Curr Opin Crit Care*. 2018;24(3):151.
9. Nishijima I, Oyadomari S, Maedomari S, et al. Use of a modified early warning score system to reduce the rate of in-hospital cardiac arrest. *J Intensive Care*. 2016;4(1):12. doi:10.1186/s40560-016-0134-7
10. Franklin C, Mathew J. Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Crit Care Med*. 1994;22(2):244-247.
11. Schein RMH, Hazday N, Pena M, Ruben BH, Sprung CL. Clinical Antecedents to In-Hospital Cardiopulmonary Arrest. *Chest*. 1990;98(6):1388-1392. doi:10.1378/chest.98.6.1388
12. Rich K. Inhospital cardiac arrest: pre-event variables and nursing response. *Clin Nurse Spec*. 1999;13(3):147-153.
13. Kause J, Smith G, Prytherch D, Parr M, Flabouris A, Hillman K. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom—the ACADEMIA study. *Resuscitation*. 2004;62(3):275-282.
14. Bedell SE, Deitz DC, Leeman D, Delbanco TL. Incidence and characteristics of preventable latrogenic cardiac arrests. *Jama*. 1991;265(21):2815-2820.
15. Wang AY, Fang CC, Chen SC, Tsai SH, Kao WF. Periarrest Modified Early Warning Score (MEWS) predicts the outcome of in-hospital cardiac arrest. *J Formos Med Assoc*. 2016;115(2):76-82. doi:10.1016/j.jfma.2015.10.016
16. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *Qjm*. 2001;94(10):521-526.

17. Morgan R, Williams F, Wright M. An early warning scoring system for detecting developing critical illness. *Clin Intensive Care*. 1997;8(2):100.
18. Churpek MM, Yuen TC, Huber MT, Park SY, Hall JB, Edelson DP. Predicting cardiac arrest on the wards: a nested case-control study. *Chest*. 2012;141(5):1170-1176.
19. Alam N, Vegting IL, Houben E, et al. Exploring the performance of the National Early Warning Score (NEWS) in a European emergency department. *Resuscitation*. 2015;90:111-115. doi:10.1016/j.resuscitation.2015.02.011
20. Almutary A, Althunayyan S, Alenazi K, et al. National Early Warning Score (NEWS) as Prognostic Triage Tool for Septic Patients. *Infect Drug Resist*. 2020;13:3843-3851. doi:10.2147/IDR.S275390
21. Spångfors M, Molt M, Samuelson K. In-hospital cardiac arrest and preceding National Early Warning Score (NEWS): A retrospective case-control study. *Clin Med Lond Engl*. 2020;20(1):55-60. doi:10.7861/clinmed.2019-0137
22. RCP. Royal College of physicians. National early warning score (news) 2: standardising the assessment of acute-illness severity in the NHS. Updated report of a working Party. In: RCP London; 2017.
23. Kimblad H, Marklund J, Riva G, Rawshani A, Lauridsen KG, Djärv T. Adult cardiac arrest in the emergency department – A Swedish cohort study. *Resuscitation*. 2022;175:105-112. doi:10.1016/j.resuscitation.2022.03.015
24. El Naqa I, Murphy MJ. What Is Machine Learning? In: El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing; 2015:3-11. doi:10.1007/978-3-319-18305-3_1
25. Mahesh B. Machine learning algorithms-a review. *Int J Sci Res IJSRInternet*. 2020;9(1):381-386.
26. Lu TC, Wang CH, Chou FY, et al. Machine learning to predict in-hospital cardiac arrest from patients presenting to the emergency department. *Intern Emerg Med*. 2023;18(2):595-605. doi:10.1007/s11739-022-03143-1
27. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using time series analysis to predict cardiac arrest in a pediatric intensive care unit. *Pediatr Crit Care Med J Soc Crit Care Med World Fed Pediatr Intensive Crit Care Soc*. 2015;16(9):e332.
28. Javan SL, Sepehri MM, Javan ML, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. *Comput Methods Programs Biomed*. 2019;178:47-58.
29. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44(2):368.
30. Wu TT, Lin XQ, Mu Y, Li H, Guo YS. Machine learning for early prediction of in-hospital cardiac arrest in patients with acute coronary syndromes. *Clin Cardiol*. 2021;44(3):349-356. doi:10.1002/clc.23541
31. Kwon J, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc*. 2018;7(13):e008678.
32. Mendes M, Pala A. Type I error rate and power of three normality tests. *Pak J Inf Technol*. 2003;2(2):135-139.
33. du Prel JB, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. *Dtsch Arzteblatt Int*. 2010;107(19):343-348. doi:10.3238/arztebl.2010.0343

34. Nti IK, Nyarko-Boateng O, Aning J. Performance of machine learning algorithms with different K values in K-fold cross-validation. *J Inf Technol Comput Sci*. 2021;6:61-71.
35. Jung K, Bae DH, Um MJ, Kim S, Jeon S, Park D. Evaluation of nitrate load estimations using neural networks and canonical correlation analysis with k-fold cross-validation. *Sustainability*. 2020;12(1):400.
36. Wayahdi MR, Syahputra D, Ginting SHN. Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification. *INFOKUM*. 2020;9(1, Desember):1-6.
37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
38. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:4768-4777.
39. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front Public Health*. 2017;5:307. doi:10.3389/fpubh.2017.00307
40. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. *Mach Learn*. 2021;110(3):451-456. doi:10.1007/s10994-021-05964-1
41. Kleppe A. Area under the curve may hide poor generalisation to external datasets. *ESMO Open*. 2022;7(2):100429. doi:10.1016/j.esmoop.2022.100429
42. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, eds. *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg; 2013:451-466.
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(null):2825-2830.
44. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed*. 2022;214:106584. doi:10.1016/j.cmpb.2021.106584
45. Mangalathu S, Hwang SH, Jeon JS. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng Struct*. 2020;219:110927. doi:10.1016/j.engstruct.2020.110927
46. Mitchell OJL, Edelson DP, Abella BS. Predicting cardiac arrest in the emergency department. *J Am Coll Emerg Physicians Open*. 2020;1(4):321-326. doi:10.1002/emp2.12015
47. Trzeciak S, Rivers EP. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. *Emerg Med J*. 2003;20(5):402-405.
48. Johnson KD, Winkelman C, Burant CJ, Dolansky M, Totten V. The factors that affect the frequency of vital sign monitoring in the emergency department. *J Emerg Nurs*. 2014;40(1):27-35.
49. Armstrong B, Walthall H, Clancy M, Mullee M, Simpson H. Recording of vital signs in a district general hospital emergency department. *Emerg Med J EMJ*. 2008;25(12):799-802. doi:10.1136/emj.2007.052951
50. Horwitz LI, Green J, Bradley EH. US emergency department performance on wait time and length of visit. *Ann Emerg Med*. 2010;55(2):133-141. doi:10.1016/j.annemergmed.2009.07.023