Predicting Patient Admission at Emergency Department Triage with Deep Super Learner Ensembles of Machine Learning Models

Tyler Benson[1]

[1]**Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing**

ABSTRACT

Objective
To predict likelihood of hospital admission during emergency department triage using dynamic triage data as it becomes available.

Background
Emergency department overcrowding is a relevant problem when its byproducts include reduced access to timely, quality clinical care, increased boarding times, and greater risk for patients and providers. Patient admission prediction models offer apparent solutions to this problem through educated resource allocation of inpatient units and accelerated decision-making during triage. However, developing a model to predict patient admission that reduces over-triaged patients while keeping under-triaged patients to a minimum remains difficult. Also, with a purpose of admission prediction being to reduce time spent during triage, it becomes necessary to develop models that can handle missing data – i.e. missing vitals upon initial patient arrival – to provide likelihood estimations during each stage of ED encounters. We hyperparameter tuned 8 individual machine learning algorithms, used a super learner (SL) to ensemble and find the best performing ensemble of models, then trained an assortment of models using the best SL model with combinations of possible missing variables removed in order to create a framework that can predict likelihood of patient admission dynamically, as each data becomes available.

Methods
The triage data, including vitals and other commonly accessible data points, of 222,260 patient encounters from a professional emergency department were recorded with 217,766 available after cleaning. The dataset was split and 43,553 (20%) patients were held to be used in a test set. Using a randomly sampled 50,000 patients in training, through the optuna framework, hyperparameter tuning was performed on eight machine learning algorithms (Random Forest, XGBoost, LGBM, Logistic Regression, Decision Tree, KNearestNeighbors, AdaBoostClassifier, ExtraTreesClassifier) with three fold cross-validation for scoring. Then, using the optimized parameters for each model, we concatenated each combination of the independent models in a SL ensemble to be tested with the highest three fold cross-validated scoring SL ensemble selected. This model was fit on varying subsets of the n=174,213 train data and scored on the test data to observe the optimal amount of data to feed the model. We then trained the highest scoring SL ensemble on the optimal amount of data with combinations of potentially missing variables removed in order to create a cluster of models able to handle variations in data availability. Finally, we evaluate the feature importances using the eli5 package's permutation importance evaluation.

Results
The overall hospital admission risk was 24.8% and decreased by triage level: ESI-1 93.5%, ESI-2 53.4%, ESI-3 25.0%, ESI-4 1.7%, and ESI-5 0.4%. A LogisticRegression baseline model, with hyperparameter tuning and trained on 50,000 patients, obtained an AUROC = 0.832 and AUPRC = 0.622. The OPSL model, the super learner model that was optimized using AUPRC as its cross-validation metric, obtained a superior score in AUROC = 0.8588 and the second best AUPRC = 0.6793 by 0.0002. It is for this reason that the hyperparameters of the OPSL model were selected as the production model and to be expanded. This model is a 7 layered ensemble with LGBMClassifier, XGBClassifier, MLPClassifier, GradientBoostingClassifier, RidgeClassifier, LogisticRegression, and Perceptron base-model predictions stacked and fed to a the meta learner, MLPClassifier. A proportion of 0.85 (n=149,260) of the training dataset was observed to be a reasonable subset of training data at which additional training data does not augment performance. Expansion of the model to handle missing variables resulted in the least knowledgeable (the model trained on data with the most amount of missing variables) model dropping 0.03 in AUROC performance compared to the fully-knowledgeable model. The ESI level was the most valuable variable per eli5's evaluation.

Conclusions
Compared to a baseline LR as well as optimized machine learning algorithms used independently, super learner ensembles of optimized machine learning models demonstrated superior performance during prediction of patient admission. Additionally, SL models trained on combinations of missing variables from patient data did not incur significant loss, suggesting that patient admission analysis can provide relevant admission likelihoods upon immediate data arrival, even if it is incomplete. These conclusions show relevance, as early recognition of a patient's likely disposition offers potential efficiencies linked to patient safety, including improving patient throughput and reducing boarding in the ED.

Keywords: Admission, Emergency department, Triage, Prediction, Super learner, Deep learning, Decision support

BACKGROUND

While most emergency department (ED) visits end in discharge, EDs constitute more than half of all patient admissions – a percentage that continues to have an upward trend [1]. ED encounters start by sorting patients by acuity in order to prioritize individuals requiring urgent medical intervention. This "triage" is typically performed by a member of the nursing staff and makes use of a small amount of available patient characteristics including but not limited to demographics, chief complaint, vital signs, and arrival mode. Thereafter, a medical care provider assesses a patient to determine their disposition.

Recent growth in available personal and medical data offers a unique opportunity for clinical optimization through application of "electronic health data and analytics, in conjunction with principles of epidemiology, health services research, and biostatistics" [2]. Data-driven research has seen success in many of its clinical applications: identification and prediction of drug effects and interactions [3], prediction of sepsis [4], predicting in-hospital mortality [5], predicting 30-day unplanned readmission [5] to provide a few examples. Although predictive algorithms cannot eliminate medical uncertainty, one such benefit is improved allocation and preparation through early-warning systems [6]. Another potential benefit to evidence-based medical predictions is efficiency and accuracy in clinical decision-making. Machine learning models have shown promise in optimizing triage processes through patient admission prediction during real-world application [7, 8].

Emergency department overcrowding is correlated with an increase in inpatient mortality and a decrease in health care quality [9, 10]. With crowding, although the time-sensitivity of critically ill patients is known, protocolized care initiation decreases and time to specifically critical therapies has been shown to increase [11]. Additionally, boarding times – in which a patient is held waiting after triage as a result of currently insufficient inpatient resources – increase as overcrowding saturates an ED. Similar to overcrowding, increased boarding time holds negative consequences and is associated with increased rates of mortality [12]. Also, capital utilized to care for boarding patients can be considered lost, as it doesn't qualify as patient services – thus higher operational revenues can be achieved through a reduction in boarding time [13]. One such solution to ED overcrowding is an improvement in information coordination between EDs and inpatient units, in that effective communication can improve quality of care in general hospitals [14] and optimize patient throughput processes [15].

With regard to improved communication leading to ample preemptive resource allocation, simple predictive models (linear regression, Naive Bayes) have been shown to improve bed management and patient flow [16, 17]. From the perspective of patient care in the ED setting, a patient's likelihood of admission assists a nurse / physician's intuitive determination of severity and can immediately communicate severity to a number of downstream decisions such as bed placement or potential emergency intervention [18]. In addition to Naive Bayes, Logistic Regression is common in classification prediction and has seen success when applied to patient admission [19, 20, 21, 22, 8]. However, simple logistic regressions are most often trained on hospital-specific data, with unique features unique to each hospital. This most apparently limits the scalability of predictive models as they become specialized to specific establishments. To respond to this, the application of more advanced machine learning algorithms is considered. In specific, deep learning models are known for their ability to handle these unique, often messy, datasets with flexibility [5, 23]. Machine learning has been applied to patient admission prediction and appears to offer efficient communicative operability and optimizations to patient throughput flow.

Deep learning is a powerful machine learning method that extracts interpretations layer by layer, allowing lower level features and patterns to define the interpretation of higher level features [24]. Different to that of the popular deep learning concept of neural networks, the layered interpretations can alternatively be defined by base learners in which each layer involves use of an independently trained machine learning algorithm concatenated in an ensemble. To summarize, logistic regressions are applied in hospital-specific and demographic-specific modeling, deep learning is applied for scalability and flexibility, and simple ensemble systems are used to improve performance [8]. Excitingly, the challenge of predicting patient

admission has not seen the use of super learner (SL) techniques [25]. Super learning is a system that finds the optimal combination of diverse base learning algorithms to be used in a deep learning ensemble model. [26, 27]. Super learning is not a neophyte in the clinical scene, however, and has seen application to data sets related to healthcare, yielding a superior performance compared to individual base learner algorithms [28]. Finally, we performed optimized hyperparameter tuning on the individual algorithms prior to the super learning's optimized ensembling analysis – a technique that may help optimize performance [29].

Another problem with patient admission prediction is held in organizational differences, in that each independent hospital may record different variables during emergency department encounters. A proposed solution involves exploring the application of models that make use of only a few highly informative predictive variables that are common among hospitals [7]. Similarly, not all of these required, predictive variables are immediately available – i.e. a full set of vitals. The challenge of missing data is common within machine learning and has many approaches to making predictions / training models on incomplete data. Some examples include complete case analysis (only patients where all variables are present are included), simple imputation (replace with mean, median, or mode), and regression imputation (include inherent relationships between variables in order to predict missing values) [30]. Complete case analysis is applicable when missing data is not overwhelming and doesn't hold inherent information in its absence; however, it doesn't possess relevance during real-world application, as it is impractical to leave out a patient's prediction until all the necessary data is acquired [31]. Rather than impute missing variables in given patient encounter, we take an altered case analysis approach when developing models for production use. That is, we choose to generate models trained on data with combinations of potentially missing data columns removed. In this, whereas imputation via regression or simple methods wields a confidence in the imputed value that can be deceptive, this technique allows a predicted likelihood of admission with relative confidence levels during every stage of a patient encounter.

An accurate prediction of inpatient admission, made early during the ED visit, could allow an earlier start to the bed allocation process and thus reduce boarding times. In contrast to common approaches, a deep learning super learner ensemble may provide the flexibility required in predicting patient admission effectively and with augmented confidence.

METHODS

Study design and setting

Retrospective data were obtained from the emergency department of Eisenhower Health with a time-frame of from 2017 to 2020. With processing and cleaning, the sample size is 222,260 patients. ["The represented EDs include a level I trauma center with an annual census of approximately 85,000 patients, a community hospital-based department with an annual census of approximately 75,000 patients, and a suburban, free-standing department with an annual census of approximately 30,000 patients"] ← TODO: spiel that describes Eisenhower. This ED utilizes the Epic EHR and makes use of the Emergency Severity Index (ESI) for triage.

Data Collection and processing

13 variables were recorded for each patient visit. The first instance of recorded vital signs was saved – temperature, heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, oxygen saturation. Additionally, the mode-of-arrival, ESI triage category, chief complaints, sex, age, month, and year were recorded. Using the Pandas data manipulation package [32], a series of processing steps were taken to clean and prepare the data for modeling:

- Chief complaints were binary encoded, resulting in 689 out of all patients) columns. Additionally, the number of chief complaints that each patient had was derived and appended as a feature.
- Mode-of-arrival was binary encoded to be no ambulance (0) or ambulance (1).
- Age categories of pediatric, adult, geriatric 65-80, and geriatric 80+ were derived from age and then one-hot encoded, resulting in 4 additional features.

- Intuitively manifested vital thresholds were derived and appended as additional binary features. This includes non pediatric temperatures of more than 104, non pediatric systolic blood pressure less than 80, non-pediatric respiratory rates of more than 40, and oxygen saturation of less than 85%
- Similarly, the categorical data of sex, month, and year were replaced by their encodings, adding 3 features for sex (male, female, unknown), 12 for month, and 3 for year.
- After this cleaning is performed, the patients that had missing variables were dropped, leaving us with 217,766 patients in total.

Response Variable

Extraneous, less-clear disposition outcomes were binned. Patients binned as admitted were those who left against medical advice (AMA), expired, were transferred, were sent to the Cath Lab / OR / Specialty Department, required observation, or were admitted. Those binned as discharged were patients who eloped, left without being seen (LWBS), or were discharged.

Model fitting and evaluation

That patient data were shuffled and 43,553 patients were withheld as a test set to be used solely in final analysis and 50,000 patients were chosen as the train sample-size.

Random Forest (RF), Logistic Regression (LR), Decision Tree, KNearestNeighbors (KNN), AdaBoost Classifier (ABC), ExtraTrees Classifier (ET) (all from scikit-learn [33]), XGBoost (XGB) [34], and LightGradientBoostingMachine (LGBM) [35], were each put into an Optuna test suite with variable parameters [36]. As per Optuna's framework, the highest scoring model for the given scoring metric is considered the optimized model and its parameters are saved. In order to bypass data leakage, at every step of Optuna's optimization, three fold cross validation on the train set is performed to produce scores. The evaluation metrics used in this study are Receiver Operating Characteristic area under the curve (ROCAUC) and Precision-Recall area under the curve (PRAUC).

Subsequently, a super learner ensemble framework was developed (adapted from Jason Brownlee's logic [27]) to be used with scikit-learn and Optuna. Optuna creates a study suite to evaluate the overall score for each combination of parameters – in the case of a SL ensemble, combinations of baselearner models – in order to make educated guesses for optimal combinations that maximize the score. After the SL study suite is finished and the optimal makeup of basemodel layers are found, its predicted probabilities for the test set are obtained and its calibration is observed through scikit-learn's calibration curve [33].

This calibrated, optimized super learner is fit on variable training subset proportions 0.15, 0.25, 0.4, 0.5, 0.7, 0.8, 0.9, and 1 out of the total train set n=174,213.

This ideal amount of training data is used in refitting the model on data with combinations of potentially missing values removed. The potentially missing variables in this study were considered to be: blood pressure, heart rate, temperature, oxygen saturation, respiratory rate, age, and ambulance. Because a given patient could have any amount of these variables missing, this becomes a 7 choose i combination problem with i ranging from 1 to 7 (7 choose 0 is the already fit and saved base case of all variables), and results in 127 additionally generated and fit models.

Finally, the variables of importance are determined and displayed via the eli5 python package [37] so as to convey the reasoning behind the model's decisions.

RESULTS & DISCUSSION

Characteristics of Study Samples

A total of 217,766 ED visits were available for analysis after filtering for exclusion criteria, with [some %] of the samples excluded due to missing values. The overall hospital admission risk was 24.8% and decreased by triage level: ESI-1 93.5%, ESI-2 53.4%, ESI-3 25.0%, ESI-4 1.7%, and ESI-5 0.4%. Additional characteristics of the study samples are presented in Table 1. The distribution of time data is attached in the appendix in Table 6.

The disposition's were binned to be admitted and discharged. There are studies utilizing the approach done here regarding disposition binning [38] as well as those who drop less clear dispositions [18] that show the success of either alternative. This study binned extraneous outcomes with the idea that their disposition can be retrospectively determined. That is, if a patient expires, it is apparent that they should have been admitted and thus the data are treated accordingly.

Table 1. Summary of the variables present in the data. The counts and normalized counts (in %) of admitted, discharged, and all patients are shown for each categorical variable. For continuous variables, the median and inner quartile range (IQR) are shown.

	Discharge	**Admitted**	**Overall**
n (%)	163784 (75.2)	53982 (24.8)	217766
Chief Complaint (top 15)			
ABDOMINAL PAIN	21156 (74.4)	7283 (25.6)	28439 (11.16)
CHEST PAIN	8674 (54.0)	7397 (46.0)	16071 (6.31)
SHORTNESS OF BREATH	5528 (47.5)	6110 (52.5)	11638 (4.57)
FALL	7622 (73.5)	2753 (26.5)	10375 (4.07)
FEVER	7520 (76.7)	2287 (23.3)	9807 (3.85)
COUGH	6375 (88.3)	844 (11.7)	7219 (2.83)
BACK PAIN	6093 (86.0)	992 (14.0)	7085 (2.78)
DIZZINESS	5497 (77.7)	1579 (22.3)	7076 (2.78)
VOMITING	5426 (76.8)	1643 (23.2)	7069 (2.77)
WEAKNESS - GENERALIZED	3429 (49.6)	3490 (50.4)	6919 (2.72)
HEADACHE	5720 (89.6)	664 (10.4)	6384 (2.51)
FLANK PAIN	4054 (86.0)	661 (14.0)	4715 (1.85)
SYNCOPE	2290 (61.7)	1423 (38.3)	3713 (1.46)
ALTERED MENTAL STATUS	1388 (38.4)	2229 (61.6)	3617 (1.42)
OTHER	2710 (80.9)	638 (19.1)	3348 (1.31)
Number of Chief Complaints			
0	633 (48.8)	664 (51.2)	1297 (0.65)
1	129237 (74.8)	43610 (25.2)	172847 (86.06)
2	18635 (69.8)	8065 (30.2)	26700 (13.29)
3	3 (100.0)	0 (0.0)	3 (0.0)
ESI Level			
1	21 (6.5)	304 (93.5)	325 (0.16)
2	22294 (46.6)	25538 (53.4)	49944 (22.93)
3	77118 (75.0)	25665 (25.0)	102783 (51.17)
4	47123 (98.3)	824 (1.7)	47947 (23.87)
5	1952 (99.6)	8 (0.4)	1960 (0.98)
Ambulance arrival			
No	128231 (79.9)	32167 (20.1)	160398 (79.86)
Yes	20277 (50.1)	20172 (49.9)	40449 (20.14)
Gender			
Female	82472 (76.6)	25154 (23.4)	107626 (53.59)
Male	66018 (70.8)	27176 (29.2)	93194 (46.4)
Unknown	18 (66.7)	9 (33.3)	27 (0.01)
Median of Continuous Variables (IQR)			
Age (years)	50.9 (29.5-70.5)	70.8 (56.0-81.1)	57.4 (34.0-74.5)
Temp (F*)	98.2 (98.0-98.5)	98.2 (97.9-98.7)	98.2 (97.9-98.6)
Heart Rate (bpm)	85.0 (76.0-94.0)	89.0 (78.0-96.0)	86.0 (77.0-95.0)
Respiratory Rate	18.0 (18.0-20.0)	20.0 (18.0-22.0)	18.0 (18.0-20.0)
Oxygen Saturation (%)	99.0 (98.0-99.0)	99.0 (98.0-99.0)	99.0 (98.0-99.0)
Systemic Blood Pressure	141.0 (126.0-159.0)	148.0 (129.0-168.0)	143.0 (127.0-161.0)
Diastolic Blood Pressure	85.0 (76.0-95.0)	84.0 (72.0-95.0)	85.0 (75.0-95.0)

Table 1 shows the top 15 most popular chief complaints. Our cleaned dataset consisted of 706 columns – 689 of which were for chief complaints after one-hot encoding. The disproportionate amount of columns for chief complaints comes with disadvantages. As some models sample fractions of features to train 'trees' on, one-hot encoding this cardinal feature will result in bias towards the one-hot encoded features as they will get selected more [39]. Also, shown in Table 1's number of chief complaints, any given patient will have 0's (meaning that they didn't express that specific chief complaint) for 687 or more of

their chief complaint columns. This can slow down learning significantly. A solution to the disadvantages of one-hot encoding would be to bin uncommon complaints as 'Other'. Doing this will reduce model dimensionality; however, within the medical context, efficient training times was not a priority (especially if it meant sacrificing the prediction confidence on rare complaints) and thus no binning was used.

[I don't exactly know where to put this paragraph, because it mentions the final model (that I haven't talked about yet)] There are alternative techniques to encode cardinal categorical variables that can reduce dimensionality – with each their own disadvantages. For the sake of empirical support, an evaluation was performed to determine the optimal encoding technique. Following the logic outlined in Viacheslav Prokopev's Kaggle Kernel, the AUROC performance of the final super learner model was obtained for the encoding techniques of one-hot encoding, label encoding, frequency encoding, and mean encoding. The definitions, advantages and disadvantages, and relevance of each of these techniques are described in his Kernel [39]. The results of this study are attached in the Appendix as Table 7 and Figure 4. As expected, the one-hot encoding technique took longer to train, reaching its optimal performance on the test set after 500 iterations. Additionally, the one-hot encoding technique yielded the best AUROC score.

It is worth noting that there are variables that may be available in the ED that this dataset does not contain. For example, lab results, radiology, or even early diagnoses have been used in alternative predictive models [40]. In terms of machine learning, as deep learning algorithms are currently the preferred approach to speech recognition and computer vision, their application is immediately exciting within the interpretation of these free-text physician notes / lab results [24, 41]. Similarly, this challenge of text mining (gathering quantitative information from text) within patient admission prediction has also seen the use of support vector machines [42]. However, these variables are only available after the patient has received significant workup – which could be hours after patient arrival – and, as this study intends the model for use in predictions in the emergency department as early as possible, these variables are not included.

Individual and Super Learner Model Optimization

The randomly sampled 50,000 patient initial parameter tuning was performed and scores were saved. These data were randomly sampled in order to be rid of temporal ordering and allow for a more diverse snapshot of categorical values, that is, using the first 50,000 patients without shuffling excludes data from the 2020 year. For each model, the area under the receiver operating characteristics curve (AUROC) and area under the precision-recall curve (AUPRC) are obtained from the held-out test set. The super learner models obtained scores superior to their optimized basemodel makeups in AUROC and AUPRC (Table 2).

The basemodels of XGBoost (XGB), LightGBM (LGBM), AdaBoostClassifier (ABC), RandomForest (RF), LogisticRegression (LR), DecisionTree (DT), ExtraTrees (ET), and KNearestNeighbors (KNN) were chosen to be optimized because they are common models in classification applications. Table 3, showing the basemodels selected to be in the optimized super learners, includes models without hyperparameter optimization: during super learner optimization, additional models that have an ability to output prediction probabilities were included as potential basemodels.

These performance metrics were chosen specifically because they don't require a single threshold to be chosen in evaluation. That is, our model outputs a prediction likelihood as a percentage and it would be necessary to set a threshold when evaluating metrics such as accuracy and F score. Because our model is intended for use as a clinical decision support tool, and will show percentage likelihood of admission rather than a binary prediction, the metrics that evaluate model performance over all possible thresholds are used. These include but are not limited to AUROC and AUPRC. AUROC is a popular metric for classification problems, summing the area under the curve of false positive rate vs true positive rate. AUPRC is a metric concerned only with the performance on the positive class, summing the area under the curve of precision vs recall. It is considered more reliable on imbalanced data sets [43] (such is ED admission data).

It can be seen in Table 2 that individual machine learning algorithms that are optimized on specific metrics don't always score the highest on their respective optimized metric. This is likely due to the inherent randomness used in hyperparameter tuning. Another reasonable explanation holds that the optuna framework brought the models to local maxima in their potential scores and deemed them as optimized when other, more optimal hyperparameter makeups existed.

Model	AUROC	AUPRC
OPSL	0.858767	0.679255
ORXGB	0.858592	0.678014
ORSL	0.858575	0.679458
OPXGB	0.857659	0.677013
ORLGBM	0.856351	0.674954
OPLGBM	0.855716	0.672190
ORABC	0.854659	0.669726
OPABC	0.854564	0.668984
OPRF	0.851316	0.664977
ORRF	0.851115	0.664478
ORLR	0.831798	0.621962
OPDT	0.828621	0.617678
ORDT	0.828621	0.617678
OPET	0.824837	0.623491
ORET	0.824837	0.623491
OPKNN	0.751398	0.521719
ORKNN	0.751398	0.521719

Table 2. Model hyperparameter optimization and super learner model scores tested on 50,000 patients are shown with values sorted to be descending by AUROC. The model names follow the key, O: optimized on, R: AUROC / P: AUPRC

SL base-models	AUROC	AUPRC
OPSL	0.858767	0.679255
LGBMClassifier	0.855962	0.674506
XGBClassifier	0.854308	0.669831
MLPClassifier	0.851274	0.666832
GradientBoostingClassifier	0.849616	0.663277
RidgeClassifier	0.846570	0.652033
LogisticRegression	0.834651	0.627932
Perceptron	0.820741	0.612998
ORSL	0.858575	0.679458
LGBMClassifier	0.856570	0.676544
MLPClassifier	0.850545	0.664913
GradientBoostingClassifier	0.849529	0.663566
SGDClassifier	0.843524	0.644850
LogisticRegression	0.831492	0.622973
Perceptron	0.831394	0.623722

Table 3. The base-model make-ups (selected via Optuna) of each optimized super learner. Tested on the held-out test set, sorted to be descending by AUROC relative to each SL.

The OPSL model, the SL model that was optimized using AUPRC as its metric, obtained the best score in AUROC and the second best AUPRC score - 0.0002 less than ORSL. It is for this reason that the hyperparameters and basemodels of the OPSL model were selected to be expanded and as the production model.

We test an assortment of machine learning models on patient encounter data to produce which algorithm may yield superior confidence in their predictions – that is, a superior AUROC or AUPRC score. The results held in Table 3 show that super learner techniques can produce a model that scores better than its optimized base-models, consistent with SL theory that states the SL will perform at least as well as its base-models [25].

Additionally, our objective is to produce a model that can output predictions in a digestible format. Validating that the optimized model produces prediction probabilities that are consistent with the information they convey is a necessity. Calibrated classifiers are probabilistic classifiers in which the predicted probability can be directly interpreted as a confidence level. For example, a patient with a 0.80 output score can be understood to have an 80% chance of being admitted. This allows the user to comprehend the confidence of each prediction and issue their own decision accordingly.

Figure 1 displays scikit-learn's calibration curve for the OPSL model. In figure generation, the OPSL model predicted probabilities for all patients in the test set and these predictions were compared to the fraction of patients with similar scores that were admitted.

The orange line of the OPSL model acts similarly to the blue, dashed line of what would be a perfectly calibrated classifier (Figure 1). Because the ratio of a patient's predicted probability of being admitted to the amount of patients with similar probabilities that were admitted is approximately 1:1, the user of this model can reliably interpret the output as an accurate representation of the chance that their patient will be admitted.

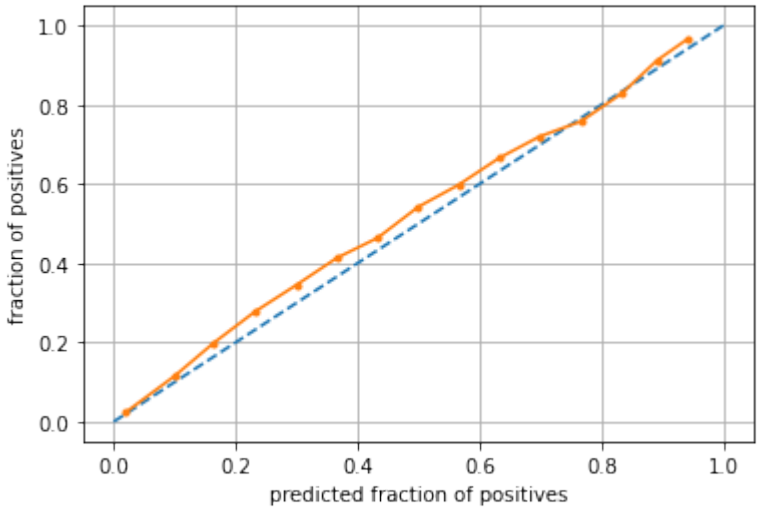

Figure 1. Calibration curve with 15 bins. The dashed blue line shows a perfectly calibrated curve with absolutely direct correlation. The super learner's predicted probabilities were placed in 15 bins and plotted as orange dots

On this note, it is undoubtedly preferable to see a predicted probability that is in the extremes of either 0 or 100%. Predictions with confidences in the extremes would be more useful to the user than would a prediction of 25% (which is the overall rate of admission to the hospital for this data set). To determine how much the preference of extreme predicted probabilities is met, Figure 2 is generated. In addition to presenting the calibration of the model by stacking the admitted vs discharged patient counts for each bin, this figure shows the amount of patients in the entire data set that received each prediction; it conveys both the rarity of a specific predicted value of a patient as well as the meaning behind the admission probability that that value holds.

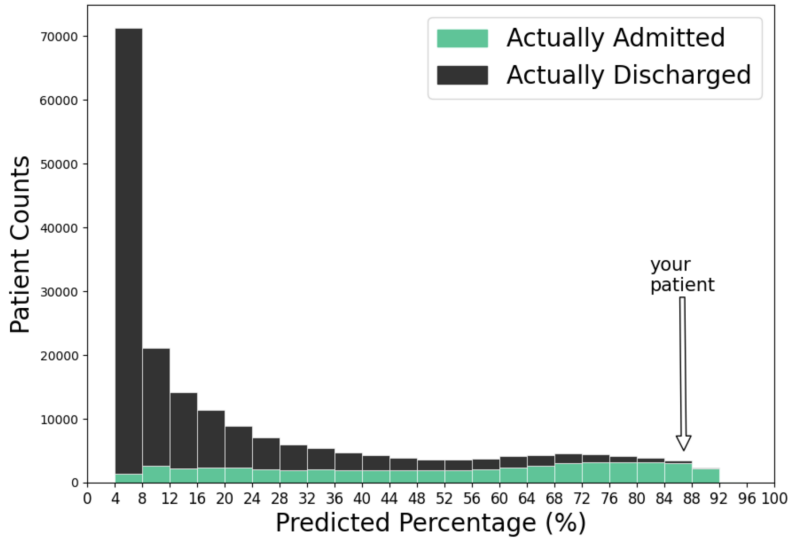

Figure 2. Histogram of predicted percentages. The green stacked bars represent the patients who were actually admitted, those in the data set with a positive target variable, and the black stacked bars represent patients who were actually discharged. A hypothetical patient with a predicted probability of admission of 87% is shown with an arrow.

Super Learner Model Optimal Subset Size

The model that scored the best during the 50k optimization study was selected as OPSL. This model was then trained on incrementally larger amounts of patients and its AUROC and AUPRC scores are shown Figure 3. This allows us to find the most efficient amount of patients to train the models during model expansion. As model expansion will train a large amount of models (+100), it is ideal to train on an efficient subset of data granted there is confidence that increasing the subset size will not yield superior scores.

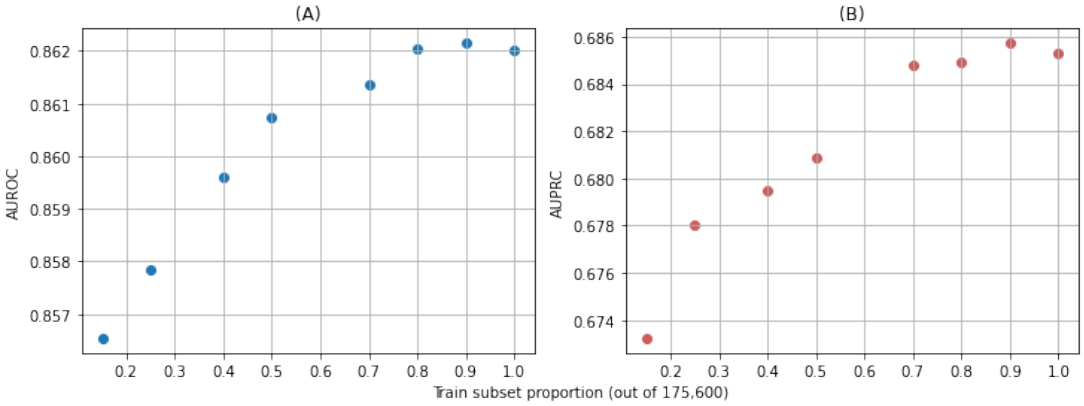

Figure 3. AUROC and AUPRC of the super learner (OPSL: SL with AUPRC as its optimization metric) trained on variable subset proportions. The model was trained on proportions 0.15, 0.25, 0.4, 0.5, 0.7, 0.8, 0.9, 1.0 of the total training size of n=174,213. The AUROC is shown in blue in Subplot (A) and the AUPRC is shown in red circle in Subplot (B)

Figure 3 shows that the super learner optimized model approaches a maximum performance as training size increases. A maximum score is expected, as all predictive models contain error given inherent noise in the data and the stochastic nature of the algorithms themselves [44]. Because the size of training data approaches a maximum score, it is inefficient to train on the total amount of data during fitting. The proportion of 0.85 (n=149,260) of the total train data set is selected to be used in fitting during model expansion.

Super Learner Model Expansion

Uniformity is not present in patient encounters in the emergency department nor is there a consistency of data collection between hospitals; a framework that allows for any combination of variables to be input into the model is necessary for its clinical applicability. This study produced a framework that involved the creation of models for every potential situation. An arbitrary thirteen model scores of this 'expansion' is shown in Figure 4 below. The rest of the models are attached in the appendix as Tables 8 and 9.

Table 4. Performance results of the evaluation of "model expansion". Defined as making and scoring a model for each potential combination of missing data as columns removed.

Removed Columns	AUROC	AUPRC
temp	0.860	0.683
HR	0.858	0.675
RR	0.858	0.676
O2	0.861	0.684
BP_sys,BP_dia	0.859	0.677
ambulance	0.858	0.678
age(and groupings),	0.851	0.666
temp,HR	0.857	0.674
temp,RR	0.857	0.674

Additionally, the generation of conglomerates of models trained on combinations of missing features did not incur significant loss, suggesting that they may provide useful insights to clinicians and inpatient units about incoming patients. The least knowledgeable model, model "temp,HR,RR,O2,ambulance,age(and groupings),BP_sys,BP_dia", obtained a 0.818 AUROC score.

An alternative approach to the necessity of a flexible model would be to use an existing algorithm that can handle missing values. For example, LGBM and XGBoost have built-in missing value handling. If missing data is present during training, these tree-based algorithms learn the optimal direction for missing data for each split (left or right). This optimal direction is then used for missing values during scoring. If no missing data is present during scoring (for a particular feature), then the majority path is followed if the value is missing [45]. These solutions imply an inherent importance in a value being "missing" – making predictions as if this is valuable information. While a patient missing a vital datapoint may have significance, because patient data is dynamic (missing values may be acquired as the patient's stay progresses), the fact that a datapoint is missing should not be used to sway admission probability.

Variables of importance

With the plethora of back-end parameter tuning and model selection, our final model may fall under the classification of a "black box" model. That is, the inner workings of the model may not be understood – inputs are given and outputs believed. In an attempt to be rid of this classification, we evaluate the feature importances that each base model in the ensemble uses in evaluation of patient data. This will allow us to reach an understanding of how our model reached its predicted probability as well as consider the most influential variables present in emergency department data.

Table 5 shows the output of the top 15 features as determined by the eli5 package [37]. The next 85 top features are attached in the appendix as Table 10. Eli5 is a machine-learning wrapper for debugging and describing "black box" estimators. We adopt the package for evaluation using "permutation importance", or "Mean Decrease Accuracy (MDA)", a technique that measures how score decreases when a feature is not available – determining the importance of each feature

Table 5. Feature importance of the top 15 features for the OPSL model as detetermined by eli5. ESI stands for emergency severity index, HR for heart rate, RR for respiratory rate, temp for temperature, CC for chief complaint, and BP_sys for systolic blood pressure.

Feature	Weight
ESI_level	0.1282 ± 0.0100
age	0.0543 ± 0.0048
HR	0.0153 ± 0.0043
ambulance	0.0086 ± 0.0024
RR	0.0073 ± 0.0013
temp	0.0050 ± 0.0014
CC_ABDOMINAL PAIN	0.0036 ± 0.0019
BP_sys	0.0029 ± 0.0025
sex_Female	0.0023 ± 0.0031
CC_CHEST PAIN	0.0022 ± 0.0010
CC_CELLULITIS	0.0018 ± 0.0005
CC_HEADACHE	0.0016 ± 0.0012
CC_ALLERGIC REACTION	0.0016 ± 0.0007
CC_PSYCHIATRIC EVALUATION	0.0014 ± 0.0004
CC_WOUND INFECTION	0.0013 ± 0.0005

There are apparent downsides to evaluating feature importance by permutation importance. For example, the chief complaint (CC) of abdominal pain is the highest valued CC; however, because it makes up 11% of the chief complaints seen in our data set (Table 1), it is obvious that the absence of this CC is a significant loss of information when compared to other (more rare) CCs and will result in a poorer performance.

Table 5 shows that ESI level is the most valuable feature. This reveals an interesting predicament. The emergency severity index acts as the indicator of triage by acuity and is defined by a triage nurse upon initial patient encounter. If our model were to uphold its reliability in the absence of ESI, its predicted probabilities may well be able to act as an independent triage tool. That is, patients that receive high probabilities of being admitted are likely to be part of a consort that are obvious in requiring immediate care; however, the hypothesis that confidence in admission is correlated to severity of emergency (while relatively observable when looking at ESI vs admission in Table 1) is not proven in this study.

These feature importances are reflected in our model expansion: the models trained without the variables of age / HR / ambulance have apparent lower scores. However, while the eli5 package tests the effect on the score when each individual feature isn't available, our model expansion tests the performance of any combination of a specific set of features that might not be available.

CONCLUSIONS

We have confirmed that the application of super learner ensembles of machine learning models can produce a robust model that scores superior to baseline models tested on the same data. Additionally, this model can be expanded to handle missing data and provide insights regarding admission during most every stage of emergency department encounters.

[Considering something like this: However, while the super learner models performed better, the difference in performance of SL to some of the single model classifiers such as XGBoost or the simpler RandomForest is minimal – that is, a RandomForest baseline model trained on all training data without tuned hyperparameters achieves a 0.85479 AUROC, a performance only 0.008 below our ORSL selected model (when trained on all training data). When considered in addition to the relevant factors of computational requirements, readability, and relative understandability [maybe expand on this point], this may suggest SL as less viable than alternative classification techniques.]

Further studies are needed to confirm best practices in conveying this information to clinicians and to the inpatient unit. Also, retrospective analysis was done on data from a pre-covid world, further studies and re-training may be needed to adapt the model to the new setting. Lastly, alternative studies have shown the success of making use of patient history and lab results as additional information; further studies applying super learner strategies to form models educated on wider sources of data may augment model confidence and applicability.

DATA AVAILABILITY

The data used in this study was derived from the electronic health records of the Eisenhower Health system and was de-identified for anonymity; however, with regard to the privacy of public health information, it is not publicly available.

APPENDIX

Table 6. Continuation of Table 1, showing the year and month categorical variables.

	Discharge	Admitted	Overall
n (%)	163784 (75.2)	53982 (24.8)	217766
Year			
2017	26006 (76.4)	8031 (23.6)	34037 (16.95)
2018	53143 (74.8)	17887 (25.2)	71030 (35.37)
2019	53374 (73.3)	19458 (26.7)	72832 (36.26)
2020	15985 (69.7)	6963 (30.3)	22948 (11.43)
Month			
1	14899 (73.8)	5285 (26.2)	20184 (10.05)
2	13320 (72.7)	5006 (27.3)	18326 (9.12)
3	13509 (73.3)	4920 (26.7)	18429 (9.18)
4	11199 (71.8)	4391 (28.2)	15590 (7.76)
5	10054 (72.6)	3785 (27.4)	13839 (6.89)
6	8264 (74.9)	2773 (25.1)	11037 (5.5)
7	12515 (76.1)	3926 (23.9)	16441 (8.19)
8	12360 (75.9)	3920 (24.1)	16280 (8.11)
9	12375 (74.6)	4210 (25.4)	16585 (8.26)
10	12722 (74.5)	4352 (25.5)	17074 (8.5)
11	13322 (73.7)	4753 (26.3)	18075 (9.0)
12	13969 (73.6)	5018 (26.4)	18987 (9.45)

Figure 4. The area under the receiver operating characteristic curve as a function of iterations is shown for various categorical encodings. The figure generation code is adapted from Viacheslav Prokopev's Kaggle Kernel [39]

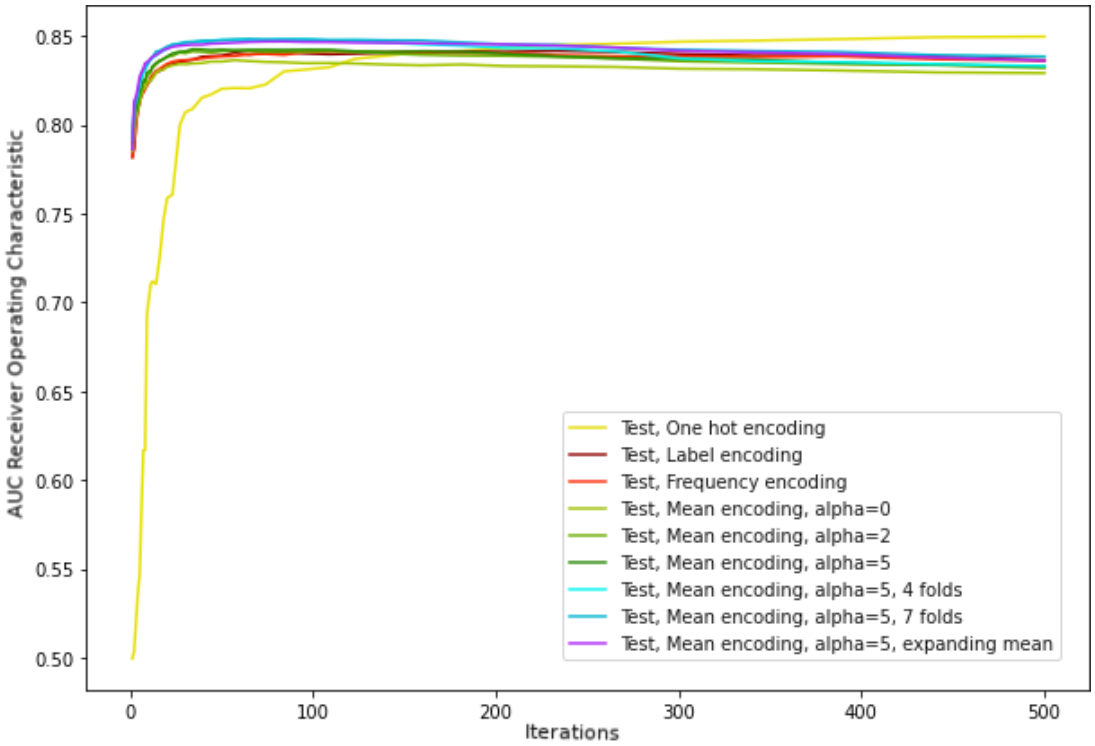

Table 7. Performance resuslts of the evaluation of different encoding techniques. The specifics of this evaluation is described in Viacheslav Prokopev's Kaggle Kernel [39]

Encoding	Train AUROC score on best iteration	Best AUROC score (test)	Best iteration (test)
One hot encoding	0.873179	0.849505	500
Label encoding	0.918103	0.841452	233
Frequency encoding	0.915529	0.841240	109
Mean encoding, alpha=0	0.929845	0.836261	57
Mean encoding, alpha=2	0.929703	0.841569	57
Mean encoding, alpha=5	0.904322	0.842399	34
Mean encoding, alpha=5, 4 folds	0.917948	0.847903	96
Mean encoding, alpha=5, 7 folds	0.918500	0.848188	65
Mean encoding, alpha=5, expanding mean	0.930003	0.846755	84

Table 8. Continuation of Table 4

Removed Columns	AUROC	AUPRC
HR,RR	0.852	0.661
HR,O2	0.857	0.674
HR,BP_sys,BP_dia	0.855	0.666
HR,ambulance	0.854	0.667
HR,age(and groupings),	0.848	0.658
RR,O2	0.857	0.673
RR,BP_sys,BP_dia	0.855	0.666
RR,ambulance	0.854	0.667
RR,age(and groupings),	0.848	0.656
O2,BP_sys,BP_dia	0.858	0.676
O2,ambulance	0.857	0.678
O2,age(and groupings),	0.850	0.663
ambulance,BP_sys,BP_dia	0.854	0.669
age(and groupings),BP_sys,BP_dia,	0.845	0.653
ambulance,age(and groupings),	0.844	0.653
temp,HR,RR	0.851	0.659
temp,HR,O2	0.856	0.672
temp,HR,BP_sys,BP_dia	0.853	0.663
temp,HR,ambulance	0.852	0.664
temp,HR,age(and groupings),	0.847	0.657
temp,RR,O2	0.856	0.672
temp,RR,BP_sys,BP_dia	0.853	0.662
temp,RR,ambulance	0.852	0.664
temp,RR,age(and groupings),	0.847	0.654
temp,O2,BP_sys,BP_dia	0.857	0.674
temp,O2,ambulance	0.856	0.674
temp,O2,age(and groupings),	0.849	0.663
temp,ambulance,BP_sys,BP_dia	0.854	0.667
temp,age(and groupings),BP_sys,BP_dia,	0.844	0.649
temp,ambulance,age(and groupings),	0.843	0.652
HR,RR,O2	0.852	0.660
HR,RR,BP_sys,BP_dia	0.849	0.652
HR,RR,ambulance	0.848	0.651
HR,RR,age(and groupings),	0.843	0.645
HR,O2,BP_sys,BP_dia	0.854	0.666
HR,O2,ambulance	0.853	0.666
HR,O2,age(and groupings),	0.846	0.656
HR,ambulance,BP_sys,BP_dia	0.850	0.656
HR,age(and groupings),BP_sys,BP_dia,	0.842	0.647
HR,ambulance,age(and groupings),	0.841	0.645
RR,O2,BP_sys,BP_dia	0.854	0.664
RR,O2,ambulance	0.853	0.665
RR,O2,age(and groupings),	0.846	0.654
RR,ambulance,BP_sys,BP_dia	0.850	0.656
RR,age(and groupings),BP_sys,BP_dia,	0.840	0.641
RR,ambulance,age(and groupings),	0.839	0.640
O2,ambulance,BP_sys,BP_dia	0.854	0.669
O2,age(and groupings),BP_sys,BP_dia,	0.843	0.650
O2,ambulance,age(and groupings),	0.842	0.649
ambulance,age(and groupings),BP_sys,BP_dia,	0.836	0.638
temp,HR,RR,O2	0.850	0.657
temp,HR,RR,BP_sys,BP_dia	0.847	0.646
temp,HR,RR,ambulance	0.845	0.645
temp,HR,RR,age(and groupings),	0.841	0.642
temp,HR,O2,BP_sys,BP_dia	0.853	0.661
temp,HR,O2,ambulance	0.851	0.662
temp,HR,O2,age(and groupings),	0.845	0.655
temp,HR,ambulance,BP_sys,BP_dia	0.849	0.654
temp,HR,age(and groupings),BP_sys,BP_dia,	0.841	0.645

Table 9. Continuation of Tables 4 and 8

Removed Columns	AUROC	AUPRC
temp,O2	0.861	0.683
temp,BP_sys,BP_dia	0.858	0.675
temp,ambulance	0.857	0.677
temp,age(and groupings),	0.851	0.667
temp,HR,ambulance,age(and groupings),	0.840	0.643
temp,RR,O2,BP_sys,BP_dia	0.853	0.662
temp,RR,O2,ambulance	0.852	0.663
temp,RR,O2,age(and groupings),	0.845	0.651
temp,RR,ambulance,BP_sys,BP_dia	0.849	0.652
temp,RR,age(and groupings),BP_sys,BP_dia,	0.839	0.637
temp,RR,ambulance,age(and groupings),	0.838	0.637
temp,O2,ambulance,BP_sys,BP_dia	0.853	0.665
temp,O2,age(and groupings),BP_sys,BP_dia,	0.841	0.648
temp,O2,ambulance,age(and groupings),	0.841	0.649
temp,ambulance,age(and groupings),BP_sys,BP_dia,	0.834	0.634
HR,RR,O2,BP_sys,BP_dia	0.848	0.650
HR,RR,O2,ambulance	0.847	0.649
HR,RR,O2,age(and groupings),	0.840	0.642
HR,RR,ambulance,BP_sys,BP_dia	0.844	0.639
HR,RR,age(and groupings),BP_sys,BP_dia,	0.836	0.631
HR,RR,ambulance,age(and groupings),	0.834	0.628
HR,O2,ambulance,BP_sys,BP_dia	0.850	0.657
HR,O2,age(and groupings),BP_sys,BP_dia,	0.840	0.644
HR,O2,ambulance,age(and groupings),	0.838	0.642
HR,ambulance,age(and groupings),BP_sys,BP_dia,	0.833	0.631
RR,O2,ambulance,BP_sys,BP_dia	0.849	0.655
RR,O2,age(and groupings),BP_sys,BP_dia,	0.837	0.638
RR,O2,ambulance,age(and groupings),	0.836	0.636
RR,ambulance,age(and groupings),BP_sys,BP_dia,	0.829	0.620
O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.832	0.633
temp,HR,RR,O2,BP_sys,BP_dia	0.845	0.645
temp,HR,RR,O2,ambulance	0.844	0.645
temp,HR,RR,O2,age(and groupings),	0.839	0.640
temp,HR,RR,ambulance,BP_sys,BP_dia	0.841	0.633
temp,HR,RR,age(and groupings),BP_sys,BP_dia,	0.834	0.627
temp,HR,RR,ambulance,age(and groupings),	0.832	0.624
temp,HR,O2,ambulance,BP_sys,BP_dia	0.848	0.652
temp,HR,O2,age(and groupings),BP_sys,BP_dia,	0.839	0.641
temp,HR,O2,ambulance,age(and groupings),	0.838	0.640
temp,HR,ambulance,age(and groupings),BP_sys,BP_dia,	0.831	0.627
temp,RR,O2,ambulance,BP_sys,BP_dia	0.847	0.651
temp,RR,O2,age(and groupings),BP_sys,BP_dia,	0.837	0.634
temp,RR,O2,ambulance,age(and groupings),	0.836	0.634
temp,RR,ambulance,age(and groupings),BP_sys,BP_dia,	0.828	0.616
temp,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.831	0.629
HR,RR,O2,ambulance,BP_sys,BP_dia	0.842	0.636
HR,RR,O2,age(and groupings),BP_sys,BP_dia,	0.833	0.627
HR,RR,O2,ambulance,age(and groupings),	0.831	0.623
HR,RR,ambulance,age(and groupings),BP_sys,BP_dia,	0.825	0.609
HR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.829	0.625
RR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.825	0.615
temp,HR,RR,O2,ambulance,BP_sys,BP_dia	0.840	0.631
temp,HR,RR,O2,age(and groupings),BP_sys,BP_dia,	0.831	0.623
temp,HR,RR,O2,ambulance,age(and groupings),	0.829	0.620
temp,HR,RR,ambulance,age(and groupings),BP_sys,BP_dia,	0.822	0.602
temp,HR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.828	0.624
temp,RR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.824	0.611
HR,RR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.820	0.603
temp,HR,RR,O2,ambulance,age(and groupings),BP_sys,BP_dia,	0.818	0.597

Table 10. Feature importance of the top **100** features for the OPSL model as detetermined by eli5.

Feature	Weight	Feature	Weight
ESI_level	0.1282 ± 0.0100	CC_SORE THROAT	0.0001 ± 0.0007
age	0.0543 ± 0.0048	CC_SWALLOWED FOREIGN BODY	0.0001 ± 0.0002
HR	0.0153 ± 0.0043	CC_ARM SWELLING	0.0001 ± 0.0000
ambulance	0.0086 ± 0.0024	CC_ITCHING	0.0001 ± 0.0001
RR	0.0073 ± 0.0013	CC_ABNORMAL LAB	0.0001 ± 0.0004
temp	0.0050 ± 0.0014	CC_WRIST INJURY	0.0001 ± 0.0002
CC_ABDOMINAL PAIN	0.0036 ± 0.0019	sex_Male	0.0001 ± 0.0004
BP_sys	0.0029 ± 0.0025	age_group_Adult	0.0001 ± 0.0002
sex_Female	0.0023 ± 0.0031	CC_ALCOHOL INTOXICATION	0.0001 ± 0.0001
CC_CHEST PAIN	0.0022 ± 0.0010	CC_HEAD LACERATION	0.0001 ± 0.0001
CC_CELLULITIS	0.0018 ± 0.0005	CC_EYE PROBLEM	0.0001 ± 0.0005
CC_HEADACHE	0.0016 ± 0.0012	CC_DIFFICULTY SWALLOWING	0.0001 ± 0.0002
CC_ALLERGIC REACTION	0.0016 ± 0.0007	CC_VASCULAR ACCESS PROBLEM	0.0001 ± 0.0000
CC_PSYCHIATRIC EVALUATION	0.0014 ± 0.0004	CC_URI	0.0001 ± 0.0003
CC_WOUND INFECTION	0.0013 ± 0.0005	age_group_Geriatric_80+	0.0001 ± 0.0001
CC_HYPERTENSION	0.0013 ± 0.0009	CC_ANKLE PAIN	0.0001 ± 0.0001
BP_dia	0.0013 ± 0.0025	CC_SUICIDE ATTEMPT	0.0001 ± 0.0002
CC_URINARY RETENTION	0.0012 ± 0.0007	CC_APHASIA	0.0001 ± 0.0001
CC_SHORTNESS OF BREATH	0.0010 ± 0.0006	CC_GROIN PAIN	0.0001 ± 0.0001
CC_DIZZINESS	0.0009 ± 0.0009	CC_FAILURE TO THRIVE	0.0001 ± 0.0001
CC_MOTOR VEHICLE CRASH	0.0008 ± 0.0007	CC_WITHDRAWAL	0.0001 ± 0.0002
CC_STROKE	0.0008 ± 0.0003	CC_PAIN	0.0001 ± 0.0002
CC_SUICIDAL	0.0008 ± 0.0005	CC_ABSCESS	0.0001 ± 0.0004
CC_HIP PAIN	0.0006 ± 0.0006	CC_ADDICTION PROBLEM	0.0001 ± 0.0001
O2	0.0005 ± 0.0007	CC_LEG PAIN	0.0001 ± 0.0001
CC_CEREBROVASCULAR ACCIDENT	0.0005 ± 0.0006	CC_NIGHT SWEATS	0.0001 ± 0.0001
CC_FEVER	0.0004 ± 0.0004	CC_RAPID HEART RATE	0.0001 ± 0.0002
CC_LACERATION	0.0004 ± 0.0005	CC_MASS	0.0001 ± 0.0001
CC_ALTERED MENTAL STATUS	0.0004 ± 0.0004	CC_BODY FLUID EXPOSURE	0.0001 ± 0.0002
CC_DIARRHEA	0.0004 ± 0.0007	CC_LOSS OF CONSCIOUSNESS	0.0001 ± 0.0001
CC_JAUNDICE	0.0003 ± 0.0001	CC_LABORING	0.0001 ± 0.0001
CC_HYPOTENSION	0.0003 ± 0.0002	CC_BACK PAIN	0.0000 ± 0.0003
CC_VAGINAL BLEEDING	0.0003 ± 0.0003	CC_POST-OP PROBLEM	0.0000 ± 0.0002
CC_PANIC ATTACK	0.0003 ± 0.0002	CC_num	0.0000 ± 0.0001
CC_ABNORMAL POTASSIUM	0.0003 ± 0.0000	CC_DENTAL PAIN	0.0000 ± 0.0001
CC_BLACK OR BLOODY STOOL	0.0003 ± 0.0002	CC_VERTIGO	0.0000 ± 0.0001
CC_RECURRENT SKIN INFECTIONS	0.0003 ± 0.0001	CC_TRANSIENT ISCHEMIC ATTACK	0.0000 ± 0.0001
CC_EYE PAIN	0.0002 ± 0.0001	CC_INSECT BITE	0.0000 ± 0.0002
CC_VOMITING BLOOD	0.0002 ± 0.0003	CC_EARACHE	0.0000 ± 0.0002
CC_HEAD INJURY	0.0002 ± 0.0002	CC_DECREASED VISUAL ACUITY	0.0000 ± 0.0000
CC_BLOOD IN URINE	0.0002 ± 0.0001	CC_PELVIC PAIN	0.0000 ± 0.0001
CC_FLANK PAIN	0.0002 ± 0.0005	CC_KNEE PAIN	0.0000 ± 0.0000
CC_COUGH	0.0002 ± 0.0004	CC_NEPHROLITHIASIS	0.0000 ± 0.0001
CC_POSSIBLE SEXUAL ASSAULT	0.0002 ± 0.0002	CC_WEAKNESS - GENERALIZED	0.0000 ± 0.0003
CC_ANXIETY	0.0002 ± 0.0011	month_9	0.0000 ± 0.0001
CC_RECTAL BLEEDING	0.0002 ± 0.0005	CC_FOREIGN BODY IN VAGINA	0.0000 ± 0.0000
CC_NOSE BLEED	0.0002 ± 0.0003	CC_FINGER LACERATION	0.0000 ± 0.0001
CC_ASSAULT VICTIM	0.0002 ± 0.0002	CC_VAGINAL BLEEDING - PREGNANT	0.0000 ± 0.0001
CC_NECK PAIN	0.0002 ± 0.0006	CC_WHEEZING	0.0000 ± 0.0001
CC_SKIN ULCER	0.0001 ± 0.0002	CC_MISCARRIAGE	0.0000 ± 0.0001

REFERENCES

[1] J. M. Pines, R. L. Mutter, and M. S. Zocchi, "Variation in emergency department admission rates across the united states," *Medical Care Research and Review*, vol. 70, no. 2, pp. 218–231, 2013.

[2] R. B. Parikh, J. S. Schwartz, and A. S. Navathe, "Beyond genes and molecules-a precision delivery initiative for precision medicine.," *The New England journal of medicine*, vol. 376, no. 17, p. 1609, 2017.

[3] N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science translational medicine*, vol. 4, no. 125, pp. 125ra31–125ra31, 2012.

[4] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," in *2019 Computing in Cardiology (CinC)*, pp. Page–1, IEEE, 2019.

[5] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[6] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *The New England journal of medicine*, vol. 376, no. 26, p. 2507, 2017.

[7] J. S. Peck, S. A. Gaehde, D. J. Nightingale, D. Y. Gelman, D. S. Huckins, M. F. Lemons, E. W. Dickson, and J. C. Benneyan, "Generalizability of a simple approach for predicting hospital admission from an emergency department," *Academic Emergency Medicine*, vol. 20, no. 11, pp. 1156–1163, 2013.

[8] Y. Barak-Corren, A. M. Fine, and B. Y. Reis, "Early prediction model of patient hospitalization from the pediatric emergency department," *Pediatrics*, vol. 139, no. 5, 2017.

[9] D. B. Richardson, "Increase in patient mortality at 10 days associated with emergency department overcrowding," *Medical journal of Australia*, vol. 184, no. 5, pp. 213–216, 2006.

[10] O. Miro, M. Antonio, S. Jimenez, A. D. De, M. Sanchez, A. Borras, and J. Millá, "Decreased health care quality associated with emergency department overcrowding.," *European journal of emergency medicine: official journal of the European Society for Emergency Medicine*, vol. 6, no. 2, pp. 105–107, 1999.

[11] 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," *Annals of emergency medicine*, vol. 61, no. 6, pp. 605–611, 2013.

[12] A. J. Singer, H. C. Thode Jr, P. Viccellio, and J. M. Pines, "The association between length of emergency department boarding and mortality," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1324–1329, 2011.

[13] T. Falvo, L. Grove, R. Stachura, D. Vega, R. Stike, M. Schlenker, and W. Zirkin, "The opportunity loss of boarding admitted patients in the emergency department," *Academic Emergency Medicine*, vol. 14, no. 4, pp. 332–337, 2007.

[14] B. Longest, "Relationships between coordination, efficiency, and quality of care in general hospitals," *Hospital Administration*, vol. 19, no. 4, p. 65, 1974.

[15] D. Golmohammadi, "Predicting hospital admissions to reduce emergency department boarding," *International Journal of Production Economics*, vol. 182, pp. 535–544, 2016.

[16] J. S. Peck, J. C. Benneyan, D. J. Nightingale, and S. A. Gaehde, "Predicting emergency department inpatient admissions to improve same-day patient flow," *Academic Emergency Medicine*, vol. 19, no. 9, pp. E1045–E1054, 2012.

[17] Y. Barak-Corren, S. H. Israelit, and B. Y. Reis, "Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow," *Emergency Medicine Journal*, vol. 34, no. 5, pp. 308–314, 2017.

[18] W. S. Hong, A. D. Haimovich, and R. A. Taylor, "Predicting hospital admission at emergency department triage using machine learning," *PLoS One*, vol. 13, no. 7, p. e0201016, 2018.

[19] C. A. Parker, N. Liu, S. X. Wu, Y. Shen, S. S. W. Lam, and M. E. H. Ong, "Predicting hospital admission at the emergency department triage: A novel prediction model," *Am J Emerg Med*, vol. 37, pp. 1498–1504, 08 2019.

[20] Y. Sun, B. H. Heng, S. Y. Tay, and E. Seow, "Predicting hospital admissions at emergency department triage using routine administrative data," *Academic Emergency Medicine*, vol. 18, no. 8, pp. 844–850, 2011.

[21] G. Abraham, G. B. Byrnes, and C. A. Bain, "Short-term forecasting of emergency inpatient flow," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 3, pp. 380–388, 2009.

[22] J. Leegon, I. Jones, K. Lanaghan, and D. Aronsky, "Predicting hospital admission for emergency department patients using a bayesian network," in *AMIA Annual Symposium Proceedings*, vol. 2005, p. 1022, American Medical Informatics Association, 2005.

[23] J. Leegon, I. Jones, K. Lanaghan, and D. Aronsky, "Predicting hospital admission in a pediatric emergency department using an artificial neural network," in *AMIA Annual Symposium Proceedings*, vol. 2006, p. 1004, American Medical Informatics Association, 2006.

[24] M. C. Chen, R. L. Ball, L. Yang, N. Moradzadeh, B. E. Chapman, D. B. Larson, C. P. Langlotz, T. J. Amrhein, and M. P. Lungren, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, 2018.

[25] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, 2007.

[26] S. Young, T. Abdou, and A. Bener, "Deep super learner: A deep ensemble for classification problems," in *Canadian Conference on Artificial Intelligence*, pp. 84–95, Springer, 2018.

[27] J. Brownlee, "How to develop super learner ensembles in python," Dec. 2019.

[28] M. F. Kabir and S. A. Ludwig, "Enhancing the performance of classification using super learning," *Data-Enabled Discovery and Applications*, vol. 3, no. 1, p. 5, 2019.

[29] J. Wong, T. Manderson, M. Abrahamowicz, D. L. Buckeridge, and R. Tamblyn, "Can hyperparameter tuning improve the performance of a super learner?: A case study," *Epidemiology (Cambridge, Mass.)*, vol. 30, no. 4, p. 521, 2019.

[30] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[31] G. J. Van der Heijden, A. R. T. Donders, T. Stijnen, and K. G. Moons, "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1102–1109, 2006.

[32] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[35] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 3146–3154, Curran Associates, Inc., 2017.

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[37] M. Korobov and K. Lopuhin, "Eli5 - a library for debugging/inspecting machine learning classifiers and explaining their predictions," Aug. 2016.

[38] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," *Crit Care*, vol. 23, p. 64, Feb 2019.

[39] V. Prokopev, "Mean (likelihood) encodings: a comprehensive study," Oct. 2018.

[40] T. Chan, G. Arendts, and M. Stevens, "Variables that predict admission to hospital from an emergency department observation unit," *Emergency medicine Australasia : EMA*, vol. 20, pp. 216–20, 07 2008.

[41] C. D. Naylor, "On the prospects for a (deep) learning health care system," *Jama*, vol. 320, no. 11, pp. 1099–1100, 2018.

[42] F. R. Lucini, F. S. Fogliatto, G. J. da Silveira, J. L. Neyeloff, M. J. Anzanello, R. S. Kuchenbecker, and B. D. Schaan, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *International journal of medical informatics*, vol. 100, pp. 1–8, 2017.

[43] J. Brownlee, "Roc curves and precision-recall curves for imbalanced classification," Jan. 2020.

[44] J. Brownlee, "How to know if your machine learning model has good performance," Apr. 2018.
[45] A. Candel, *H2O Driverless AI – an artificial intelligence (AI) platform for automatic machine learning. (documentation)*, 2017. http://docs.h2o.ai/driverless-ai/1-8-lts/docs/userguide/faq.html.