LIMITS TO PREDICTING HOSPITAL ADMISSION AT EMERGENCY DEPARTMENT TRIAGE USING MACHINE LEARNING

COS 597E FINAL PROJECT

Kathy Chen, Dale Lee, Sunnie Kim, Mengzhou Xia
{kc31, dalelee, suhk, mengzhou}@princeton.edu

December 8, 2020

1 Introduction

Emergency departments (EDs) are the largest source of hospital admissions and are increasingly being overloaded with patients [2, 4, 5]. Because longer patient wait times can significantly impact patient mortality, morbidity with readmission, and length of stay, it is very important that triage systems are reliable and accurate [2]. ED triage systems, in which health professionals assign priority levels to patients based on patient acuity, rely on health professionals' training, knowledge, and experience. Numerous works [5, 6, 7, 8, 9] have proposed improving the triaging system with statistical models, based on the idea that models can incorporate much larger volumes of triage and/or clinical historical data and, in theory, provide health professionals with more "objective" decisions.

To better understand the limits of this prediction problem, we analyze the results from a publication by Hong et al. [5], which makes the claim that machine learning models can robustly predict hospital admission at emergency department triage. Hong et al. train binary classifiers using logistic regression (LR), gradient boosting (XGBoost), and deep neural networks (DNN) on three datasets: one containing only triage data, one containing only patient history data, and one that incorporates both triage and patient history data. They evaluate the performance of these classifiers based on test AUC, sensitivity, and specificity and find that classifiers trained on the dataset containing both triage and patient history data outperform models trained on only one of triage or patient history data. The authors don't make claims about which classifier is the best (all perform similarly across all three datasets, though XGBoost is slightly better in terms of AUC), but more generally conclude that machine learning robustly predicts hospital admission at ED triage.

Our project explores several limits to prediction with respect to predicting hospital admission at ED triage. We first take a look at misclassified cases and determine which variables differed most significantly between misclassified samples and the entire test set. We find that age, number of admissions, and certain rare events (like poisoning or hypoxia) are often associated with misclassified samples. Next, we evaluate each classifier's sensitivity to data perturbations for a few selected features. Our results suggest trends in the classifier prediction for gender, ethnicity, language, race, and insurance. We then apply a method for improving model interpretability to the logistic regression and DNN models and present importance scores for the top 50 features in these two models–this extends the work by Hong et al. [5] to identify the top variables by information gain from the XGBoost classifier. We discover a set of features that appear to be most important for all three models and highlight particular features that are uniquely important to each classifier. Finally, we consider the generalizability of the classifiers to specific hospitals. We find that there is surprising variability between patient demographics between the three hospitals in the dataset used in the paper and also find a decrease in performance when retraining the XGBoost classifier only on data from two of the larger EDs and predicting on the held-out smaller ED.

While our work focuses on the limitations in the classifiers and data provided in Hong et al. [5], we believe that the conclusions we draw from studying this publication can provide general insights into the difficulties of evaluating and deploying machine learning (ML) systems in clinical settings.

This is a final project for the COS 597E Limits to Prediction seminar at Princeton University. Authors are listed in alphabetical order. Please find our code at `https://github.com/xiamengzhou/597E-Final`.

2 Related Work

2.1 Existing triage systems

Hinson et al. [4] surveys ED triage literature to assess the performance of triage systems around the world based on interrater reliability (i.e. the end result from triaging should always be the same, regardless of which nurse makes the assessment) and clinical outcomes such as mortality, critical illness, and hospitalization. They collected 32 studies evaluating outcomes in ED populations, with 20 studies reporting the sensitivity and specificity of triage designation for the outcome of inpatient hospitalization. The sensitivity of designating hospitalized patients at moderate to high levels of triage was usually high, with only 3 out of 20 studies reporting a sensitivity of less than 70%, and most studies reported high sensitivity (>90%) for identifying patients with ED mortality as high acuity at triage. However, sensitivity was low (<80%) for identifying patients with critical illness outcomes as high acuity at triage. Particularly concerning were 2 studies that reported sensitivity for identifying elderly patients that experience life-threatening events, which found that 11-23% of elderly patients were triaged to only moderate acuity. In the end, Hinton et al. found that all triage systems performed similarly, with common weaknesses identified across all systems. They emphasized that high sensitivity for triage assignment on these outcomes relies on triage systems being reliable. Rater reliability, measured with Cohen's κ statistic (ranges from 0-1, 0 being no agreement and 1 being perfect agreement), varied widely across all studies, with only 26% of studies reporting κ above 0.8.

2.2 Survey of current machine learning methods

Several studies have applied ML to predicting hospital admission, and other similar outcomes, at emergency department triage. [5, 6, 7, 8, 9] Many recent publications seem to favor the gradient boosting method for this task. Klug et al. [9] train a gradient boosting model to predict early (<2 days) and short-term mortality (2-30 days post ED registration) using ED triage data from a single hospital with AUCs of 0.962 and 0.923, respectively. Jiang et al. [7] trained 4 different machine learning classifiers (2 gradient-boosted models, logistic regression, and random forest) on triage data to predict triage levels for patients with cardiovascular disease and found that XGBoost performs slightly better than other models at this task (AUC 0.937). In a follow-up paper, Hong et al. [6] demonstrate that gradient boosting models trained on clinical data outperform logistic regression models trained on administrative data in predicting 72-hour and 9-day return to the ED.

Joseph et al. [8] train four ML models (logistic regression, 2-layer neural network, XGBoost, and a complex neural network model incorporating text data) to predict critically ill patients (mortality or ICU admission within 24 hours) at ED triage using triage data from a single care center and found the complex neural network model performed the best (AUC 0.85). They emphasize that models requiring use of structured diagnosis lists and past medical histories are unrealistic, because most ML approaches will not easily integrate with commercial EHR systems. Moreover, models will likely need to be tuned and adapted to a health system's specific population. They also note some shortcomings to their approach; for example, having abnormal vitals is a sufficient criteria for ICU admission in many hospitals, so the presence of these signs can create a self-fulfilling prophecy that artificially inflates the accuracy of ICU admission prediction, independent of illness severity.

In a review of machine learning models to aid triaging in emergency departments, Fernandes et al. [2] also note that XGBoost seems to have the highest average performance in terms of AUC for hospital admission prediction. While most methods report high AUCs, the authors point to lack of generalizability as a major limitation of these studies. Specifically, most ML models are trained on data collected from a single center or different hospitals in the same geographic region–which likely share similar working practices, data recording methods, and patient demographics. This is somewhat in contrast to Joseph et al.'s [8] argument, which suggests that models should not be expected to generalize to other hospitals at all. Inspired by this discussion, in Section 4.4, we investigate the generalizability of the classifiers in Hong et al. by using data from different EDs for training and testing.

Finally, Fernandes et al. [2] highlight publications where ML models have been used and validated in ED. They find that most studies report an improvement in health professionals' decision-making to be more consistent and reliable. However, models can only improve triaging if health professionals are properly trained and receptive to using the predictions to inform their decisions. Fernandes et al. point out that more robust metrics are needed to evaluate ML-based triage models if they are going to be widely deployed in practice: more studies are needed to verify that models outperform health professionals' triage predictions and quantify how incorporating the model can improve care or use of resources. Joseph et al. [8] also mention that metrics reported for models should match clinicians' needs; for example, most clinicians prefer to use diagnostic thresholds maximizing test sensitivity at triage at the expense of reduced specificity.

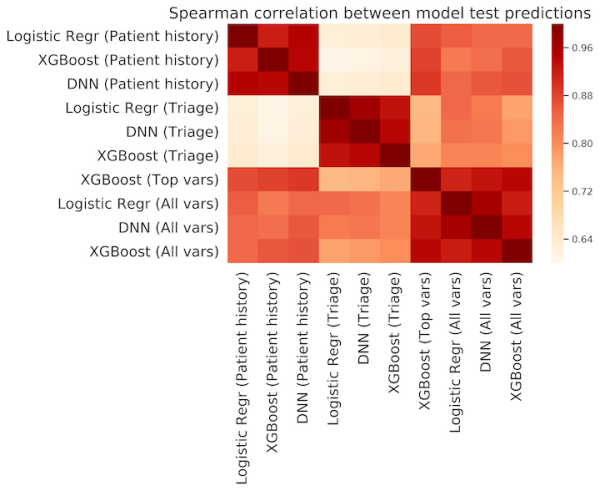

Figure 1: Spearman's correlation between the models' test predictions shows that models trained on the same set of data are most correlated with each other. Predictions made by the model trained on only the top variables (by XGBoost information gain) are most similar to the predictions made by models trained on all variables. **Number of variables in each dataset**: 222 in triage, 753 in patient history, 966 in all variables, 62 in top variables. Note that the triage dataset and patient history dataset share demographic variables, so they don't add up to 962. The paper specifies 972 variables in total but for this analysis we combine some in the same category that were one-hot encoded, see Setup (section 3).

3 Setup

Hong et al. [5] have publicly released their R code and de-identified dataset used for their analysis.[1] They describe in their publication that they obtained retrospective ED data from March 2013 to July 2017. Three EDs are represented in the data: a level trauma center, a community hospital-based department, and a suburban, free-standing department. The de-identified dataset consists of 560,486 rows (samples) and 972 columns (variables). The response variable is the patient's disposition, encoded as a binary variable where 1 denotes hospital admission and 0 denotes discharge. The dataset is randomly split into a test set of 56,000 samples and a training set of 504,486 samples.

Overall, the authors' code was easy to understand and run. We conducted most of our analysis in R and Python after running the authors' R scripts to save the trained models, datasets, and predictions. For the model interpretability analysis in Section 4.3, we exported the models trained in R to Python, as our chosen method of analysis [15] is supported in Python.

Some of the variables in the datasets are kept separate and binarized (one-hot encoded). In Section 4.1, we collapse these variables for our analysis and store the original categorical values. We also discarded a variable that could not be matched back to the full list of variables provided in [5] Supplementary Table 1. We used interpretable (easy-to-read) names for variables visualized in our figures whenever possible. However, for most variables (particularly measurement-based variables, e.g. those from blood tests), Hong et al. did not provide cleaned labels.

4 Results

In this report, we use the datasets and models reported in Hong et al. [5] as an opportunity to analyze and understand key concerns in the adoption of ML-based models in ED triage systems. Our analysis of misclassified samples shows a potential argument against including sparse variables in these classifiers; it also highlights variables available at triage that may not be sufficient to accurately predict the hospital admission for a patient. Our evaluation of model sensitivity by data perturbation reveals key variables that would likely contribute to a model's poor performance if it was adapted in a new hospital (i.e. greater sensitivity can result in poorer generalizability). We also examine what variables are most informative to the different classifiers to identify the similarities and differences between what each model learns from the data. Finally, since previous work has pointed out the question of model generalizability to different hospital settings, we also present results from retraining classifiers on data held-out for a specific ED.

[1] `https://github.com/yaleemmlc/admissionprediction`

4.1 Do misclassified samples have something in common?

We are interested in understanding what variables in the datasets the different models find important for prediction. To this end, we first focus the classifiers' predictions on the test set. We can gauge how similar the models are by looking at the correlation between their predictions (Figure 1). Consistent with the results of Hong et al. [5], models trained on the same dataset are most correlated with each other–though it's worth noting that all models' predictions are fairly highly correlated with each other (note the lowest correlation between models is between models trained on one of patient history or triage data, and even they are strongly correlated, $\rho > 0.6$). The "top variables" model, a model trained using the top 62 variables with the greatest information gain based on the "all variables" XGBoost model, is highly correlated with the models trained on all of the data (both triage and patient history data). This corroborates Hong et al.'s claim that the top variables model is trained on variables highly predictive of hospital admission.

We then identified the samples that were misclassified by each classifier and determined what variables in the data differed most significantly between the misclassified samples and all test samples. Because the classifiers output

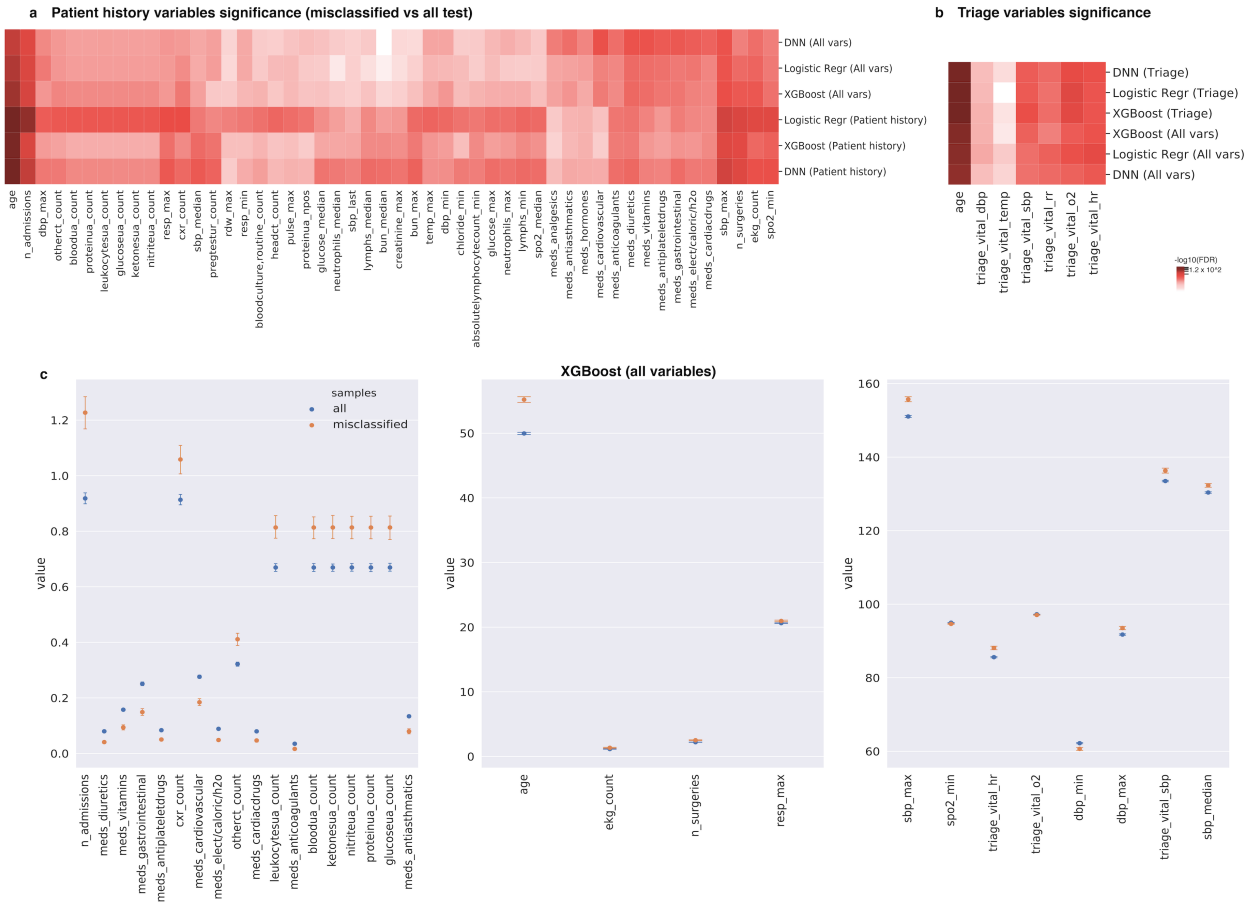

Figure 2: Numeric variables show significant differences in distribution when comparing misclassified samples to all test samples. **a, b.** Heatmaps (hierarchically clustered) to show the -$\log_{10}$(FDR-adjusted p-value) for top significant variables for patient history (only top 30) and triage data. Significance is comparable across classifiers. Colorbar is log-scaled. Note we only show classifiers containing the relevant features. We also excluded the top variables XGBoost from the visualization, but the results are similar to all variables XGBoost. **c.** The mean and 95% confidence intervals for the 30 most significant variables ($p < 0.001$) for the all variables XGBoost classifier by Wilcoxon rank sum test, FDR-adjusted using Benjamini–Hochberg over all models and numeric variables. Because the variables are measured at very different scales, we separate them into 3 subplots to better visualize the difference in mean between misclassified and all test samples. The most significant variables tend to be shared across different classifiers (for the relevant dataset on which the classifier was trained), so the XGBoost all-variables results is shown as a representative result. Plots for all other models are generated and provided in `misclassified_samples_analysis.tar.gz` as a supplement.

probabilistic predictions, we binarized the predictions based on Hong et al.'s pre-specified thresholds, which maintain a specificity rate of 0.85 for each classifier [5].

The variables are almost evenly divided into categorical (including binary) and numeric (including continuous) variables. To compare the misclassified samples to all test samples, we apply the Wilcoxon rank sum test [11] for each numeric variable and compute the log-fold change enrichment for each categorical variable.

For numeric variables, we find **age** to be the most significant variable in the dataset (Figure 2a, b), where misclassified samples tend to represent older patients (average of 55 years) compared to the entire test set (average of 50 years) as well as patients with high **n_surgeries** which represents the number of surgeries and procedures within the past year. Another variable that makes it hard to predict hospital admission is **n_admissions**, which represents the number of in-patient admissions within the past year; again, misclassified samples tend to have a higher average (>1.25) compared to all samples (<1) (Figure 2c). These three variables are immediately available in triage data and can indicate that more information is needed to make a "better" prediction for patients with these traits (older, more surgeries, more in-patient admissions this past year). There are many measurement-based variables from patient history and triage (i.e. patient vitals) that also show significant differences, but they are more difficult for us, as non-experts, to interpret.

For categorical variables, we found that the log fold-change (logFC) enrichment is generally weak, with no variable exceeding an enrichment greater than 2 (Figure 3). The top enriched variables are variables such as intrauterine hypoxia and poisoning, many of which are sparse (<1% of patients have measurements for these in the dataset) and seemingly rare events.

Figure 3: Log-fold change enrichment for categorical and binary variables, filtering only to those with enrichment above 1 (limited to top 12 if there were many). Log-fold change is computed for each variable based on the proportion of misclassified samples in a variable's class (i.e. poisoning = 1) compared to the proportion of all test samples in a variable's class. No categorical variables in the "top variables" XGBoost classifier had an enrichment above 1.

Overall, it seems likely that many misclassified samples occur because of patients having record of a particularly sparse variable; that is, these are patients that experience an illness or event rarely observed in the rest of the population. It is more surprising that we can see a clear difference in misclassification for traits such as patient age, since it's known that patients over age 60 have higher rates of hospital admission, and this percentage increases with age [1]. We speculate that age might be correlated with other variables that we highlighted, or these "less-predictable" older patients might have limited patient histories, or there is unmeasured information that can cause these samples to be anomalous.

4.2 How sensitive is each classifier to data perturbations?

Next, we study how each classifier changes its output prediction probability in response to different data perturbations. Specifically, we study the LR, XGBoost, and DNN models trained on the full set of variables.

We conduct a simple data perturbation analysis, where at inference time we change the value of one variable in the test data and observe the change in the trained model's prediction probability. We were inspired to do a data perturbation

analysis by recent perturbation-based model interpretability methods [3, 10, 12]. However, we note that the method we use, also known as one-at-a-time sensitivity analysis [14], is a much simpler method that has a long history in the field of statistics.

As it is infeasible to study all 972 variables in the dataset, we selected 10 demographic and hospital usage variables or those that have high mean information gain, calculated by Hong et al. with 100 training iterations of the full XGBoost model. For the XGBoost model, we made data perturbations to the test data, consisting of 56,000 samples. For the LR and DNN models, we made perturbations on the imputed test data where missing values were imputed with the median [5].

The **Emergency Severity Index (ESI)** is used to stratify patients into five groups based on acuity and resource needs: 1 (resuscitation), 2 (emergent), 3 (urgent), 4 (less urgent), 5 (non-urgent). It's typically assigned by a triage nurse who initially examines the patient. In fact, to the best of our knowledge, ESI level is the only variable in the data that reflects a medical professional's opinion, aside from the final hospital admission decision. We expect the ESI level to be a strong predictor of the final hospital admission decision. Indeed, perturbing the ESI level quite significantly changes the models' output prediction probability. The mean change is as big as -0.642 in the LR model when lowering ESI level from 1 (resuscitation) to 5 (non-urgent). These results are consistent with our intuition that high ESI level would indicate higher probability of hospital admission.

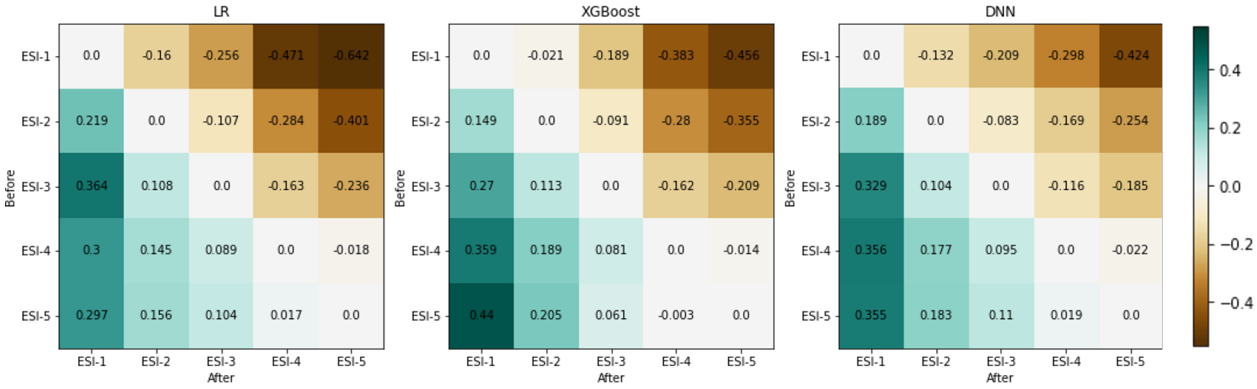

Figure 4: Mean change in the models' output predicted probabilities when we change the value of the **ESI level** variable, while keeping the values of all other variables the same.

Gender is encoded as a binary variable (Female/Male). It's unclear if this variable denotes sex at birth, gender identity, or gender expression, although we would like to note that none of these is binary. For samples with gender label Female, we change the gender label to Male and calculate the mean change in probability scores. The results are 0.018 for LR, 0.013 for XGBoost, and 0.016 for DNN. For the opposite direction where we change samples with gender label Male to Female, the mean change is -0.018 for LR, -0.012 for XGBoost, and -0.016 for DNN. These results suggest that all three models tend to predict higher probability of hospital admission for Male patients than Female patients, when all other variables are the same.

Ethnicity is encoded as a categorical variable with four fields: Hispanic or Latino, Non-Hispanic, Patient-refused, and Unknown. As there are not many Patient-refused and Unknown samples, we only study Hispanic and Non-Hispanic. Changing Hispanic to Non-Hispanic led to a mean probability change of 0.007 for LR, 0.001 for XGBoost, and 0.005 for DNN. Changing Non-Hispanic to Hispanic resulted in -0.008 for LR, -0.002 for XGBoost, and -0.007 for DNN. These results suggest that all three models tend to predict higher probability of hospital admission for Non-Hispanic patients than Hispanic patients, when all other variables are the same. However, the difference is small.

Language is encoded as a binary variable (English/Other). Changing English to Other led to a mean probability change of -0.001 for LR, 0.000 for XGBoost, and 0.001 for DNN. Changing Other to English resulted in 0.001 for LR, 0.000 for XGBoost, and -0.001 for DNN. The mean score change is too small to draw a conclusion.

Race is encoded as a categorical variable with five fields: American Indian or Alaska.Native (A), Asian (S), Black or African American (B), Native Hawaiian or Other Pacific Islander (P), and White or Caucasian (W). For each model, we measure the mean probability change for all possible combinations. While XGBoost has close to zero change for all pairs, LR and DNN models have bigger changes, especially when changing from or to Native Hawaiin or Other Pacifiic Islander, suggesting that they may be more sensitive to race than XGBoost.

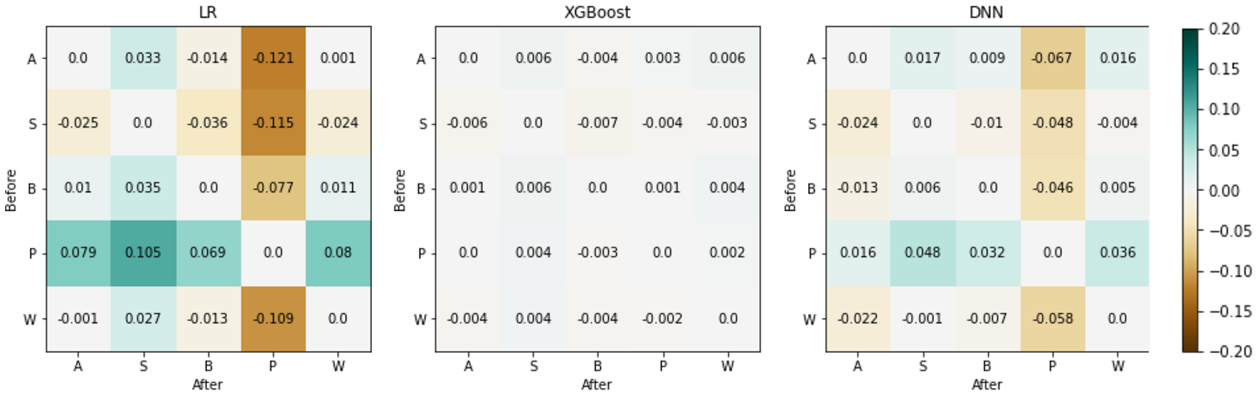

Figure 5: Mean change in the models' output predicted probabilities when we change the value of the **race** variable, while keeping the values of all other variables the same.

Insurance is encoded as a categorical variable with five fields: Commercial, Medicaid, Medicare, Other, and Self-pay. For each model, we measure the mean probability change for all possible combinations. For all models, we observe that the mean probability change from and to Self-pay is huge, as big as 0.8 in some cases. That is, all three models by and large output a higher prediction probability for hospital admission for patients with self-pay insurance. This result is in contradiction with that of Ruger et al. [13] who found that under-insured (Self-pay and Medicaid) patients were less likely to be admitted to the hospital, compared to other insured groups. Unfortunately, the current analysis doesn't provide a reason for this trend. Conducting an in-depth analysis of the insurance variable might be an interesting direction for future work.

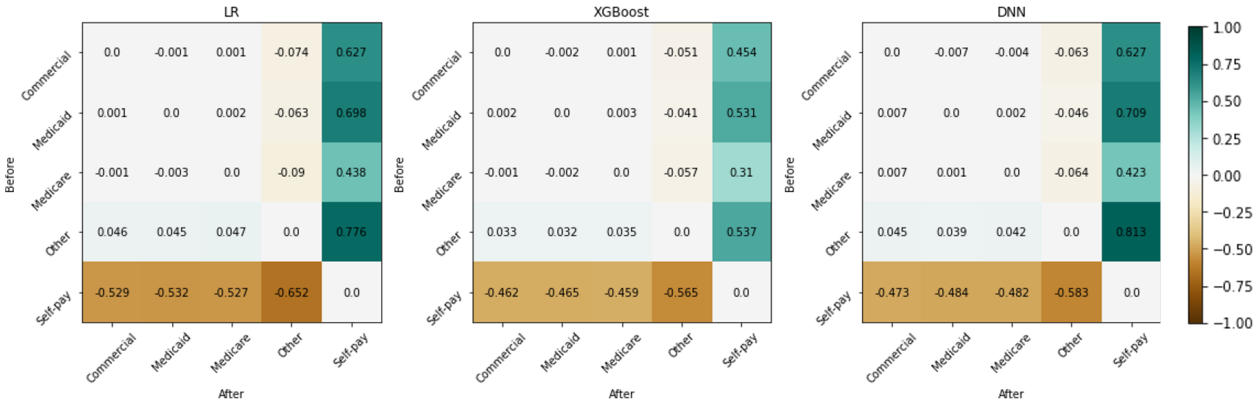

Figure 6: Mean change in the models' output predicted probabilities when we change the value of the **insurance** variable, while keeping the values of all other variables the same.

Previous disposition records the disposition of the patient's last visit to the ED: Admit, Against medical advice (AMA), Discharge, and Eloped. The direction of probability change is largely consistent among the three classifiers. If the patient was previously admitted to the hospital, the models tend to output a higher probability for hospital admission, although the mean change is less than 0.1 in magnitude for all combinations.

N_edvisits, **n_admissions**, and **n_surgeries** records the number of ED visits within the past year, the number of in-patient admissions within the past year, and the number of surgeries and procedures within the past year, respectively. As these are nonnegative integer variables, we perturb them by adding or subtracting an equally spaced sequence of integers. For example, **n_edvisits** ranges from 0 to 374, with more than 90% patients with less than 10 visits. We perturb this variable by adding 20 numbers from -100 to 100 with an interval of 10. The numbers are appropriately normalized for LR and DNN models' test data, which is why their x-axis range is different from that of the XGBoost model. Overall, we see that the models output lower prediction probabilities when the number of previous ED visits is high, possibly because patients with a high number of ED visits may have a tendency to visit EDs with less urgent symptoms. Among the three models, the mean probability change is greatest in the LR model, suggesting that the LR model is more sensitive to perturbations to the **n_edvisits** variable. The mean probability change was smaller for

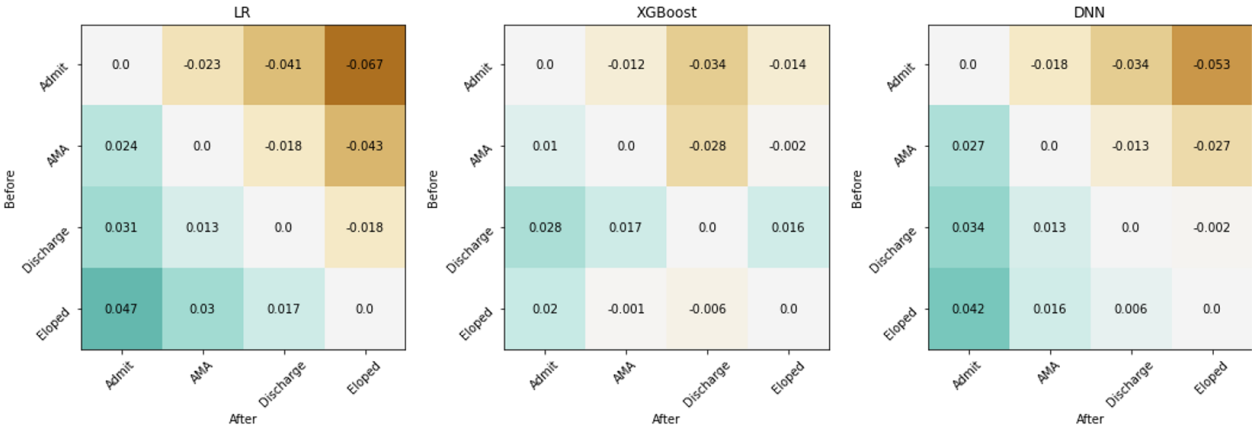

Figure 7: Mean change in the models' output predicted probabilities when we change the value of the **previous disposition** variable, while keeping the values of all other variables the same.

n_admissions and **n_surgeries**, but all three models showed a consistent direction of change: the higher the number of previous admissions or surgeries, the higher the predicted probability is for hospital admission.

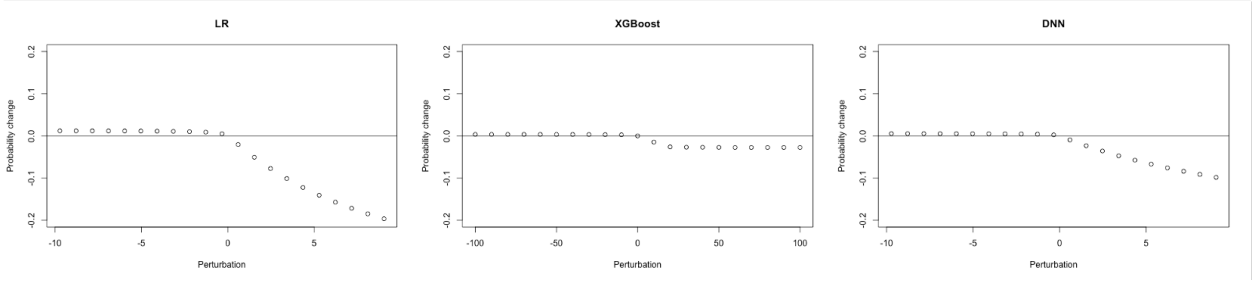

Figure 8: Mean change in the models' output predicted probabilities when we change the value of the **n_edvisits** variable, while keeping the values of all other variables the same.

While the data perturbation analysis of this section allows us to simply analyze each variable's effect on a model's prediction, it has several limitations. Mainly, it does not provide a comprehensive picture as it only shows the change in prediction instead of explaining the prediction. With this analysis, we can discover variables that exhibit big changes in prediction probabilities and hypothesize that they are important for prediction, but it is difficult to make conclusions about variables that incur small prediction probability changes. Hence in the next section, we adopt a more sophisticated model interpretability method to better understand how complex models make their predictions.

4.3 How can we interpret the decisions of each classifier?

Although all three models exhibit great performance on predicting hospital admission from ED triage, systematic analyses are lacking on revealing what basis they make predictions. In this section, we apply a model interpretability method, called Integrated Gradients [15] to probe the attribution of input features with regard to the predictions. Note that this method is only applicable on the LR and DNN models as they require analysis of gradients. Integrated Gradients achieves two desirable characteristics: sensitivity and implementation invariance. The axiom of sensitivity means that if the input and the baseline differ in one feature but have different predictions, then this feature should be assigned a non-zero contribution. The axiom of implementation invariance is less irrelevant in our analysis since the networks are quite simple.

Integrated Gradients: We first lay out the gist of the Integrated Gradients method. Formally, suppose we have a neural network $F : \mathbb{R}^n \to \mathbb{R}$, where n is the number of input variables to the neural network. We define $x \in \mathbb{R}^n$ as the input variables, $x' \in \mathbb{R}^n$ as the baseline variables. An attribution of the prediction at feature x relative to the baseline x' is a vector $A_F(x, x') = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, where a_i is the attribution of x_i to the prediction $F(x)$. The method of Integrated Gradients, as its name reflects, accumulates gradients along the path from the baseline feature x' to the input feature x to capture additional information presented in the input feature from the baseline variables.

Baseline: A key step of applying this method is to select a baseline, which is supposed to have a near-zero score. As adversarial examples would be very much likely to have the same effect, we would additionally like the baseline to convey a complete absence of signal. Usually, for image data, an image with all black pixels is a good baseline as it does not provide any signals for recognition. For textual data, an input with all zero embeddings has the same effect. Selecting a baseline feature is a bit more challenging for structured data in our case. Following common practice, we use the medium value of all variables as our baseline feature, the same as what is done in the imputation step for filling the NaNs.

Models: We apply the Integrated Gradients method to both the LR model and the DNN model. The LR model can be regarded as a one layer neural network with a sigmoid function as its activation function. Specifically, we examine the following two models: 1) A full DNN model with a 92.26 AUC, 2) A full LR model with a 91.17 AUC. In order to more flexibly adopt interpretation methods on the neural models, we decide to import the models from R to Python. Specifically, we save the weights of the models from R and build PyTorch models with the saved weights. Loading the models in PyTorch allows us to get access to the gradients of the output with regard to the input.

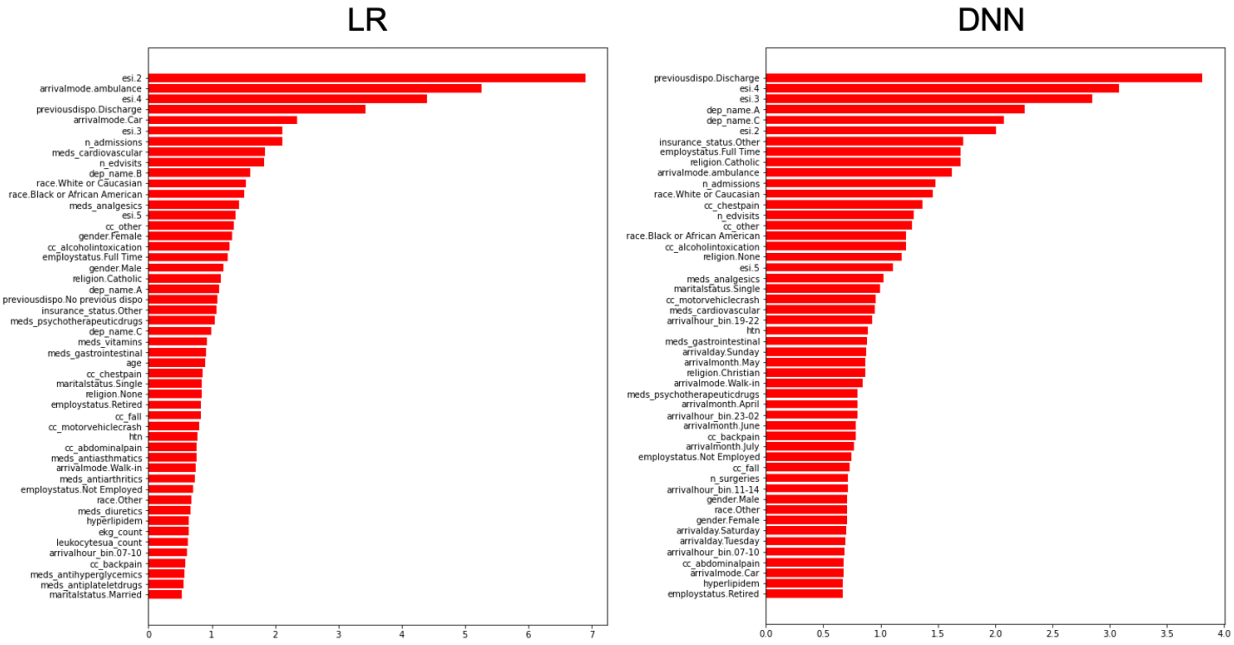

Figure 9: Importance of variables in the LR and DNN models, quantified by the Integrated Gradients method [15].

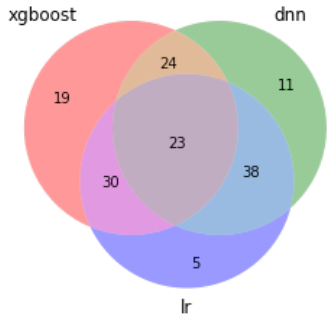

Figure 10: Venn diagram of variables that play an important role in the LR, XGBoost, and DNN models trained on the full dataset.

Results: We plot the importance scores of the top 50 variables from these two models in Figure 9, and show a Venn diagram for overlapped features of these three models in Figure 10. The Venn diagram shows that a significant number of top variables are shared between any two models, even between the XGBoost model and the other two, which adopt different methods for calculating the importance.

More specifically, there are a set of variables that consistently plays an important role in all of the three models, namely the ESI level, the arrival mode, department name, past medical history and past ED history, and employment status. On the other hand, the three models have emphasis on different sets of variables. The LR model puts more weight on the arrival mode and past medical records while the DNN model puts more weight on past medical and emergency records.

One interesting extension would be to remove the variables with the highest importance scores e.g., ESI level, to see if the model would possibly shift emphasis to other variables, which could also potentially be useful for breaking confounding effects between variables.

4.4 Are the trained classifiers generalizable to hospitals not represented in the training data?

Finally, we investigate whether hospital admission prediction results vary across hospitals and, subsequently, how generalizable the classifiers trained on data from certain hospitals are to other hospitals. We hypothesized that individual hospitals—depending on their specialists, size, and business of the ER—could have different hospital admission policies after triage. The hospitals involved in this study, though in the same hospital system, are significantly different. The largest ED is the level 1 trauma center, with 85,000 patients annually, which we refer to as ED A. The second largest ED is a community-hospital based ED, with around 70,000 patients annually, which we refer to as ED B. The last ED is a suburban, free-standing ED with only 30,000 patients, which we refer to as ED C.

Additionally, because all three of the EDs are in the same hospital system, we wondered if more advanced or well-staffed EDs would typically receive more serious patients. For example, a level 1 trauma center is a very comprehensive facility that is capable of providing complete care for every aspect of injury, while a free-standing ED does not necessarily provide the same level of complete care. Typically, ambulances will not deliver patients to free-standing EDs—patients typically transport themselves and are later driven to a hospital by ambulances if they need surgery or other more complicated procedures.

In Figure 11, we look at gender, ethnicity, and insurance type breakdown for each ED to assess if there was any significant spread, despite the EDs being relatively close to each other and being part of the same hospital system. From these graphs, we observe differences in racial breakdown and insurance type across EDs. ED C, the free-standing ED, has significantly more white patients (90% compared to 42% or 51%). We can also see that ED C has significantly more patients with commercial insurance (53% versus 26% or 34%). Though these are superficial differences in patient breakdown based on demographic and insurance information, we speculated that these differences can indicate a difference in the kind of hospital admittance procedure at each respective ED.

To evaluate the generalizability of the machine learning approach the authors took in their paper, we decided to see how well the models the authors created could predict hospital admission for specific EDs. We did this by training only on the data from two EDs (using the combined triage and patient history data), and then using a test set from the third ED. We predicted outcomes for all three EDs.

In Table 1, we compare the admit rate between the three EDs, along with the accuracy, sensitivity, specificity, and AUC of the XGBoost model in predicting hospital admissions for the ED given training data on the other two EDs. Interestingly, we find that the free-standing ED has the highest admit rate among the three EDs, despite being the smallest ED and not being attached to a hospital. When we examine the accuracy, sensitivity, specificity, and AUC, we see that the classifier performs significantly better for ED A and B than for ED C. This supports our hypothesis that the standard of care for a free-standing ED might be different from a typical trauma center or ED attached to a hospital. Importantly, there is a significant decrease in sensitivity for ED C. We see that the classifier has a sensitivity (true positive rate) of 0.63, which is much lower than the other two EDs, which have sensitivities in the low 0.90s. A lower sensitivity indicates that the classifier is less accurate in correctly predicting the patients who will be admitted

Table 1: Results of predicting hospital admission for a specific ED, using a classifier trained on data from all other EDs.

	ED Type	Num Patients	Admit Rate	Accuracy	Sensitivity	Specificity	AUC
ED A	Trauma Center	85,000	0.27	0.96	0.91	0.99	0.95
ED B	Hospital	75,000	0.29	0.97	0.94	0.99	0.96
ED C	Free-Standing	30,000	0.32	0.88	0.63	0.93	0.78

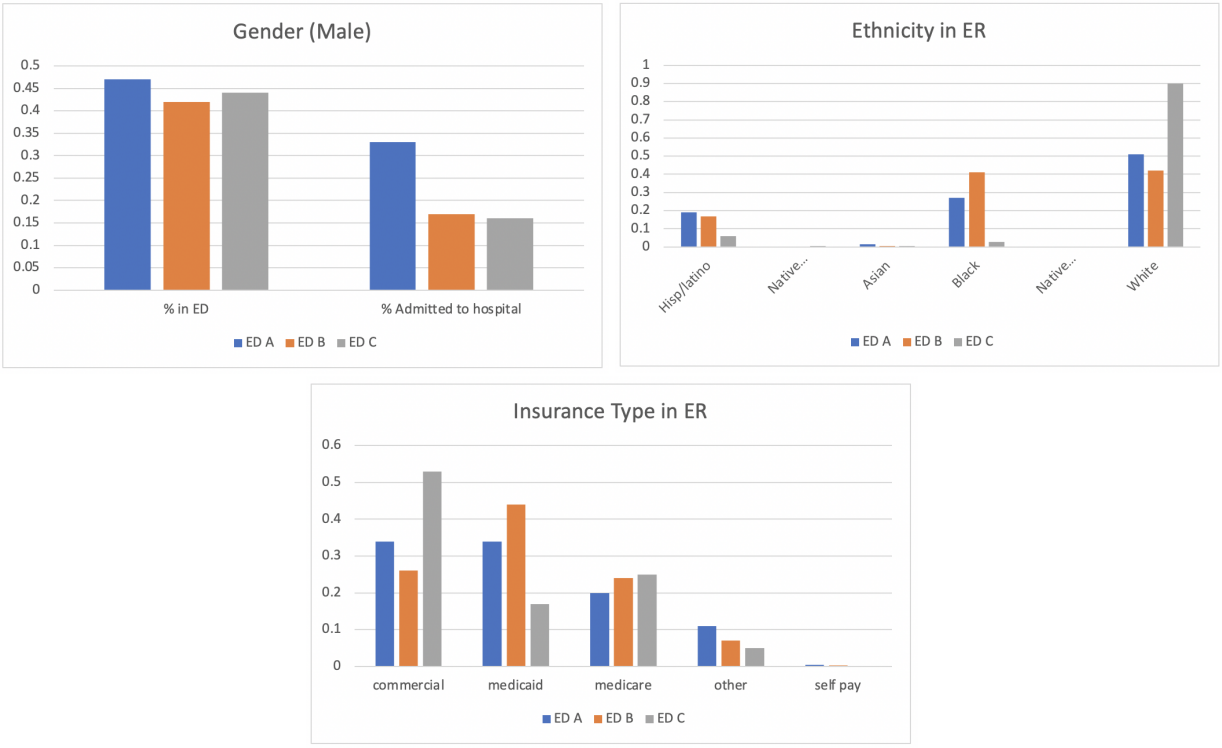

Figure 11: Gender, ethnicity, and insurance type breakdown for each of the three EDs in the dataset.

to the hospital. This kind of error is especially dangerous for medical predictions, especially if a machine learning approach like the one Hong et al. take would be used to aid making decisions on individual patients.

In Figure 12, we show the ROC curves for the three ED predictions, from which we can see that all three follow the same angled curve where the sensitivity increases very quickly and then improves very little beyond a certain threshold. We also see that ED A and B follow very similar curves, while ED C has a significantly different curve and significantly lower AUC. This supports our earlier hypothesis that ED C has a different procedure in handling hospital admittance when compared to EDs A and B.

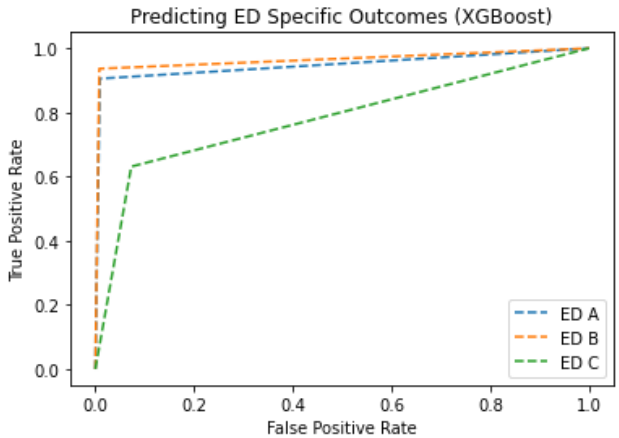

Figure 12: ROC curves for predicting hospital admisssion for an out-of-domain ED.

5 Discussion

From simple to complex, all models presented by Hong et al. all have a high test AUC of 0.87–0.92, based on which they claim that machine learning robustly predicts hospital admission at ED triage. To examine this claim and to better understand what prevents these models from being used in practice today, we conducted analyses on misclassified samples and on model sensitivity, interpretability, and generalizability.

In Section 4.1, we examined the samples misclassified by these high-performance classifiers and observed that many of them have a "rare" patient history. It is not new news that machine learning models perform worse on samples underrepresented in the training data, as they are trained to maximize accuracy on the majority of samples by default. However, misclassification, in particular false negatives, is particularly harmful in the healthcare domain, as they would deny patients from receiving the necessary care in a timely manner. Hence, it would be essential to continue to improve these models, and make specialized efforts to reduce false negatives and improve performance on underrepresented samples.

In Sections 4.2 and 4.3, we investigated how sensitive and interpretable the three types of models are. We found that all models are highly sensitive to some variables like ESI level and insurance status, changing their predicted probability and flipping the prediction in response to data perturbations. Next, we more systematically analyzed the variables and quantified their importance in the LR and DNN models, creating an analogous table to Hong et al.'s that from XGBoost. However, the insights to the models' decision rules that can be gained from these statistics are still far from sufficient and comprehensive, compared to those that can be gained from more interpretable models such as decision trees. Still, among the LR, XGBoost, and DNN models, the substantial overlap of important variables and the similarity of performance, suggest the simplest and the most interpretable LR model as an attractive option for real-world deployment in the medical domain where interpretability is highly valued.

Finally in Section 4.4, we studied the out-of-domain generalizability of the models by training and testing them with data from different hospitals. We observed that there is a significant difference in racial breakdown and insurance type between the three EDs, and that there is a small but noticeable drop in performance when we make predictions for one ED using the models trained on data from the other EDs. While expected, this performance drop opens up a number of questions on deploying these models in clinical settings. Mainly, how can we take into account the specific characteristics of each ED when training these models? How can we quantify and communicate their generalizability? How frequently should these models be updated, and how can we instill knowledge of extraordinary circumstances such as the COVID-19 pandemic into these models? These questions, as well as the earlier mentioned limitations of these models, suggest many possible directions for future work towards building truly robust and reliable models for hospital admission prediction at ED triage and likewise problems.

6 Conclusion

We present an analysis of the data and models in Hong et al. [5] and explore several limitations of predicting hospital admission using classifiers trained on emergency department triage and patient history data. We find that misclassified cases are often associated with certain rare variables that are sparsely expressed in the datasets the authors trained on, and we note specific trends in the most important features of the three classifiers. Finally, we consider the generalizability of a classifier on distinct hospitals by predicting patient outcomes in each distinct ED by training on the data from the other two EDs. We find that the three EDs have significantly different demographic and insurance breakdowns, and find that the smaller, free-standing ED is more difficult to predict using the classifier developed with data from the other two EDs. The results we've highlighted in this report demonstrate some of the complexities and limits to prediction associated with predicting patient outcomes using triage and EHR data.

References

[1] J.J. Ashman, S.M. Schappert, and L. Santo. Emergency department visits among adults aged 60 and over: United states, 2014–2017, June 2020. NCHS Data Brief No. 367.

[2] M. Fernandes, S. M. Vieira, F. Leite, C. Palos, S. Finkelstein, and J. Sousa. Clinical decision support systems for triage in the emergency department using intelligent systems: A review. *Artificial intelligence in medicine*, 102, 2020.

[3] R.C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] J.S. Hinson, D.A. Martinez, S. Cabral, K. George, M. Whalen, B. Hansoti, and S. Levin. Triage performance in emergency medicine: A systematic review. volume 74(1), pages 140–152, 2019.

[5] W.S. Hong, A.D. Haimovich, and R.A. Taylor. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, 13(7):1–13, 07 2018.

[6] W.S. Hong, A.D. Haimovich, and R.A. Taylor. Predicting 72-hour and 9-day return to the emergency department using machine learning. *JAMIA open*, 2(3):346–352, 2019.

[7] H. Jiang, H. Mao, H. Lu, P. Lin, W. Garry, H. Lu, G. Yang, T. H. Rainer, and X. Chen. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *International journal of medical informatics*, 145:104326, 2020.

[8] J. W. Joseph, E. L. Leventhal, A. V. Grossestreuer, M. L. Wong, L. J. Joseph, L. A. Nathanson, M. W. Donnino, N. Elhadad, and L. D. Sanchez. Deep-learning approaches to identify critically ill patients at emergency department triage using limited information. *Journal of the American College of Emergency Physicians Open*, 1(5):773–781, 2020.

[9] M. Klug, Y. Barash, S. Bechler, Y. S. Resheff, T. Tron, A. Ironi, S. Soffer, E. Zimlichman, and E. Klang. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: Devising a nine-point triage score. volume 35(1), pages 220–227.

[10] S.M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017.

[11] M. Neuhäuser. *Wilcoxon–Mann–Whitney Test*, pages 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[12] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[13] J. P. Ruger, C. J. Richter, and L. M. Lewis. Association between insurance status and admission rate for patients evaluated in the emergency department. *Academic emergency medicine*, 10(11), 2003.

[14] A. Saltelli, K.Chan, and E.M. Scott. *Sensitivity Analysis*. John Wiley Sons, 2000.

[15] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.