# Augmenting the Accuracy of Emergency Department Triage using Machine Learning: A Multi-Class Classification Model

Joelle Fitzgerald

## Introduction & Task

Timely and effective care of patients in the emergency department (ED) relies on cohesive workflow and optimization of resources including clinician time and attention - arguably the most critical resource in emergency care. The Centers for Medicare and Medicaid Services (CMS) measures an ED's propensity to provide safe and quality medical care by significant timestamps during a given patient's visit. Specifically, decreasing the percentage of patients who leave without being seen by a medical professional, reducing the average ED patient length of stay, and minimizing the amount of time an ED patient waits before they are admitted and transferred to another hospital unit. All of these instances in a patient's care provide significant evidence of an ED unit's efficiency and resource management and rely on timely judgement of patient acuity upon a patient's admission to the ED. To achieve and meet these CMS core measures, ED units are challenged with coordinating care and making rapid decisions from the door to time of patient discharge or admission - managing unexpected and emergent patient influxes with rationed professional attention and resource availability.

One strategy that hospitals use to prioritize and streamline safe and effective care is by triaging patients based on acuity of presenting condition. Emergency department workflow begins with an initial triage assessment - a preliminary clinical assessment performed to identify the patients who require more or less urgent care, according to the severity of their condition and before any diagnostic and therapeutic interventions. Triage is frequently performed by nursing staff, who categorize patients in order to prioritize or delay their care depending on the situation and severity of each case ranging from 1.0 (least urgent) to 5.0 (most severe/urgent). The correct assessment and triage classification of patients is associated with higher patient safety, lower morbidity and mortality. Therefore, triage decisions need to be accurate and immediate, in order to preserve clinical workflow, resources, and promote positive patient outcomes. However, elements of human bias and error during triage often results in over or under-triaging patient needs resulting in delay of care, higher medical costs, and low patient satisfaction. Removing human bias in patient acuity classification may provide an opportunity for medical staff to have a more accurate view of patient needs and preserve ED workflow, clinician/nurse time and attention, and make medical care more effective and efficient over time.

Automated systems and machine learning (ML) focused on optimizing patient acuity classification and medical triage can be used to reduce human bias and aid medical professionals in prioritizing patient needs early on in emergency care. The goal of this ML application project is to possible answer the research question: Can machine learning (ML) provide a data-driven alternative to current emergency care triage and produce an accurate patient acuity score based on minimal patient information upon registration to the Emergency Department (ED), thus reducing human-biased medical decisions and preserving clinician time, attention, and ED workflow?

## Methods

To answer this question, I will be utilizing data from the MIMIC-IV-ED dataset, intended to support data analysis in emergency care by providing a large database of admissions to an ED at a tertiary academic medical center in Boston, MA. This dataset is a module of MIMIC-IV, meaning the information contained within MIMIC-IV-ED is linkable to information in MIMIC-IV. Access to triage acuity scores (ranging 1.0-5.0, 1.0 indicating highest severity and 5.0 least emergent) and patient data such as patient arrival, admission, and discharge times, demographics (gender and race), medication information, and initial vital sign records allows for ML model to learn patterns and derive important features to develop a multi-class classification model to enhance medical triage in the ED.

Given a large dataset of over 2 million observations and multiple features, Step 1 of creating a well-developed ML application includes a brief Exploratory Data Analysis (EDA) and data pre-processing of all relevant data points/features. Exploring all possible datasets provided by MIMIC-IV and MIMIC-IV-ED, triage, ED stays, admissions, and patient datasets were chosen to merge relevant clinical and demographic information that would be applicable to initial triage assessment in the ED. From this data, a ML model can select feature importance and build an algorithm to classify any given patient into one of the five triage categories. The intended outcome of a ML medical triage model is to quickly and accurately identify patient acuity with minimal assessment information upon patient admission to the ED.

## Exploratory Data Analysis & Data Pre-Processing

In Step 1, the triage data set including initial patient vital signs (temperature, heart rate, respiration rate, oxygen saturation, systolic blood pressure, diastolic blood pressure, pain score), patient chief complaints or reason for visit, and patient acuity scores (triage score, ML model class outcome variable) were all assessed for missing and/or inaccurate values. For missing vital signs, NAs were replaced with the median value for that vital sign. Pain scores were standardized by referencing the standard scale of 1-10, 1 indicating no pain and 10 indicating severe pain - only keeping patients whose pain was recorded to be numerical 1-10. This method of reducing variability in pain scores was used in hopes to remove unintended medical bias if pain was recorded on a differing pain rating scale such as the faces scale or unable to be assessed. There were no instances of missing chief complaints. All rows that had missing acuity scores were dropped for the purpose of this project. Next, ED stays csv file was cleaned for missing values and demographic and other feature categories such as race, gender, arrival transport, and patient disposition (discharge or transfer location) were re-grouped in a concise manner in order to clean up the feature space. The patient's age was also calculated and grouped into separated age categories as well as alive or deceased status during ED visit using the patient csv. After EDA of the admissions file, it was decided that the categories of data did not add value to the set of features and was therefore left out of the model creation process.

**Feature Selection & Feature Scaling**

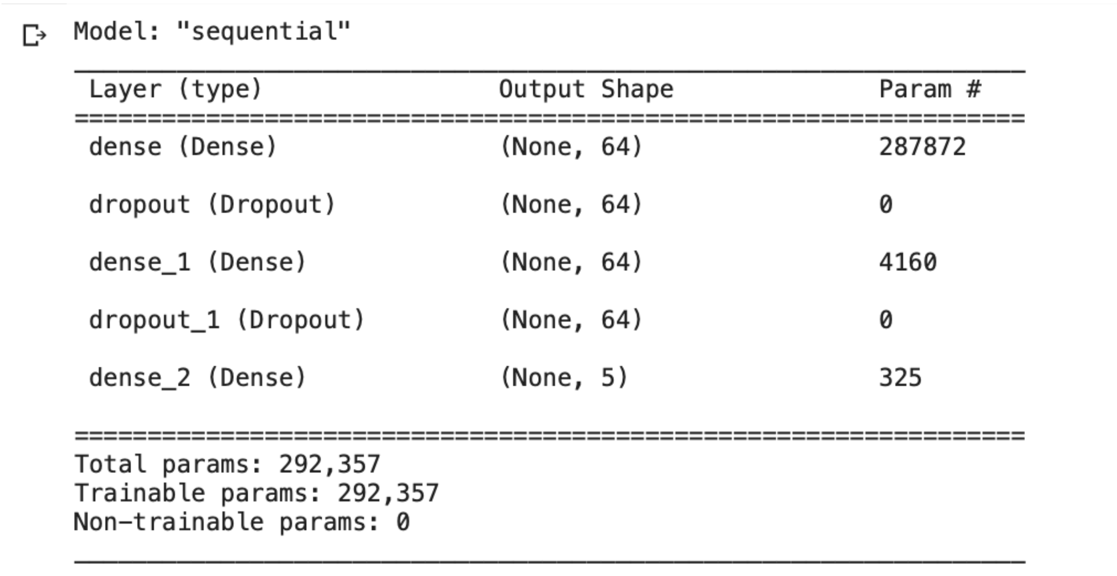
After completing the initial data pre-processing, the separate csv files were merged on subject ID to create one large dataframe of features with one outcome column - patient acuity. Categorical variables such as patient chief complaint, gender, race, arrival Transportation, disposition, age category. In order to preserve the integrity of patient chief complaint - a string of reported signs and symptoms that tell a concise story of a patient's reason for visit - each chief complaint was transformed to be lower case, unique chief complaint were found, and dummy variables were produced by one-hot-encoding each unique patient chief complaint (reason for ED admission). This process was intended to make the model more accurate in predicting patient acuity scores as a chief complaint "chest pain" may be triaged differently than a patient experiencing "chest pain, altered mental status". This distinction is important as an accurate prediction and classification between acuity scores results in preserved workflow and resources. Once dummy columns were produced for all categorical variables, the shape of the data was almost 2 million observations and around 5 thousand features. However, when assessing the distribution of acuity scores recorded in the data, class 3.0 (~60% of the data) and 2.0 (31% of the data) showed dominance while acuity class 5.0 was a minute portion of the observations occupying 0.31% of the data. In order to combat possible overfitting in the model (biasing the majority classes leading to poor performance), an undersampling approach was used to resample and balance the data. Using a random undersampling method, the majority class label frequency was computed and the ratio of majority class to minority class is defined for each label. A RandomUnderSampler object was created with a specified sampling strategy, and then applied to the 'X' dataframe and 'y' dataframe to resample the dataset. The resampled features and labels are then concatenated into a new dataframe called 'resampled_df'. By undersampling the majority class in comparison to the minority classes, the weight of each class label was evenly distributed to prevent the possibility of overfitting and increased the generalization ability of the model.

**Training & Fitting the Model**

With a more balanced outcome label distribution, the data was split into training and testing sets using an 80/20 split with ‘X’ including all available features and ‘y’ including the intended outcome column - patient acuity. The ‘y’ dataframe of patient acuity scores was one-hot-encoded to incorporate a binary multi-class outcome dataframe. Now having testing and training data, feature importance was explored on training data to find features with high importance for predicting patient acuity. Using Random Forest and XGBoost methods based on gradient boosting and using decision trees as base learners, initial patient vital signs showed high promise for predicting patient acuity. This result is important to note as patient vital signs upon admission into the ED are a primary assessment finding used to triage patients in real time. Additionally, demographic features such as gender, race (White or Black, specifically), and age (young adult) were ranked within the top 12 features, indicating demographic information play a significant role in patient acuity scoring and provide evidence that a machine learning model may aid clinicians in medical triage as more factors of a patient’s condition are accounted for instead of initial vitals and patient chief complaint. It is also important to note that chief complaints were prioritized showing higher feature importance for straightforward cheif complaints such as suicidal ideation and allergic reaction as well as those indicative of immediate care being needed such as chest pain and altered mental staus. This is a positive result of feature importance ranking as it demonstrates that the model is sifting through the priority and of each chief complaint.

Working with a multi-class classification problem, multiple methods and approaches to fitting a multi-label model were considered. The one-vs-all (or one-vs-rest) method where each classifier is trained to distinguish that class from all the other classes while training a separate binary classifier for each class, is simple but can not used because it can suffer from imbalance problems and would not be optimal for this dataset. Ensemble methods using random forests, gradient boosting machines (GBMs), and support vector machines (SVMs) and adapted for multi-class classification by modifying the underlying algorithm and combining multiple models was also considered but not selected. For the purpose of this project, a neural network model architecture was chosen using appropriate output layers including ReLU and softmax activation functions to produce multiple outputs and meaningful results.

Defining the architecture of the neural network (see Figure 1), two fully-connected dense layers of 64 units using a ReLU activation function as well as two dropout layers (0.5) in between with a final dense layer of 5 using a softmax activation function were used and ran the best metrics. The ReLU activation function was used in the dense layers to combat any vanishing gradient. The dropout layers added in between the dense layers also help to reduce overfitting by randomly dropping out a certain percentage of the input units during training. The output layer has 5 units, corresponding to the 5 possible patient acuity classes, and uses a softmax activation function to produce a probability distribution over the classes. The softmax activation function takes the exponent of each element of the input vector, making them all positive, and then normalizes them by dividing by the sum of all the exponentiated values. This normalization ensures that the output vector sums to 1 and each element is between 0 and 1, representing a probability. In neural networks, the softmax activation function is typically used in the output layer for multiclass classification tasks. The optimizer used after defining the neural network's architecture is Stochastic Gradient Descent (SGD) with a learning rate of 0.01 and Nesterov momentum. Using an SGD optimizer works by updating the model's parameters in small batches, rather than processing the entire training set at once. This ensures efficient computation of the gradients of the loss function with respect to the model's parameters. By updating the parameters based on small batches of data, SGD can escape from local minima and converge to a better global minimum. SGD also introduces some randomness into the optimization process. By randomly shuffling the training data and selecting different mini-batches for each iteration, SGD can prevent the model from overfitting to the training data.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                287872

 dropout (Dropout)           (None, 64)                0

 dense_1 (Dense)             (None, 64)                4160

 dropout_1 (Dropout)         (None, 64)                0

 dense_2 (Dense)             (None, 5)                 325

=================================================================
Total params: 292,357
Trainable params: 292,357
Non-trainable params: 0
_________________________________________________________________
```

Figure 1: Defining the neural network architecture

This architecture proved to be successful in producing a steady validation accuracy and validation loss overtime. The validation accuracy represents the percentage of correctly classified examples in the validation set, while the validation loss represents the value of the loss function on the validation set. The plot (see Figure 2) shows the values of the validation accuracy and validation loss at each epoch or iteration of the training process. From Figure 1, the validation accuracy is increasing while loss - or the difference between the predicted output and the true output for a specific input - is decreasing which means the model is fitting the data appropriately, learning patterns, and performing well.

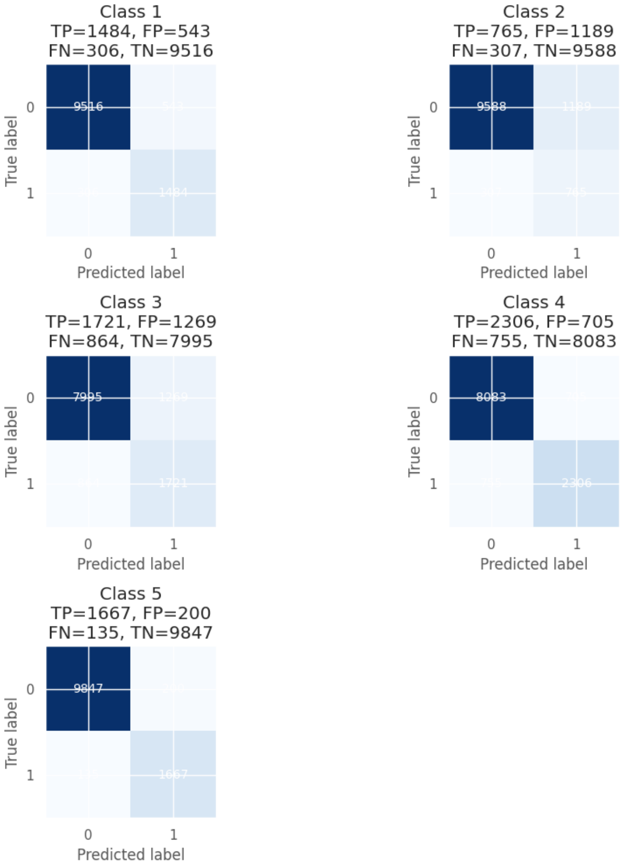


Figure 2: Training and validation accuracy and loss of defined neural network model to predict patient acuity scores

## Results

Knowing this, a classification report and model metrics were produced to understand the actual predictions of the model. A confusion matrix of true positive, true negative, false positive, and negative counts for each class was generated to understand the accuracy and error rate of the model's predictions (see Figure 3). Overall, the rate of false negatives was low compared to the other confusion matrix labels. This indicates that a small number of patients were over or under triaged. Looking at the classification report (see Figure 4), we can see that the accuracy of the model predicting class 1 and class 5 showed the most validity. This is important to note as the model predicts high risk patients who need to be seen immediately and patients who may be non urgent with a high precision and recall. While the model performed an acceptable triage classification of labels 2, 3, and 4, the variability in precision and recall may be due to the challenge of categorizing patients when small changes differentiate between middle-scale triage patient conditions and potential severity prognosis. Minute differences in a patient's vital signs or additional chief complaint detail could drastically change a triage decision and prioritize a patient higher or lower on the patient acuity score scale.

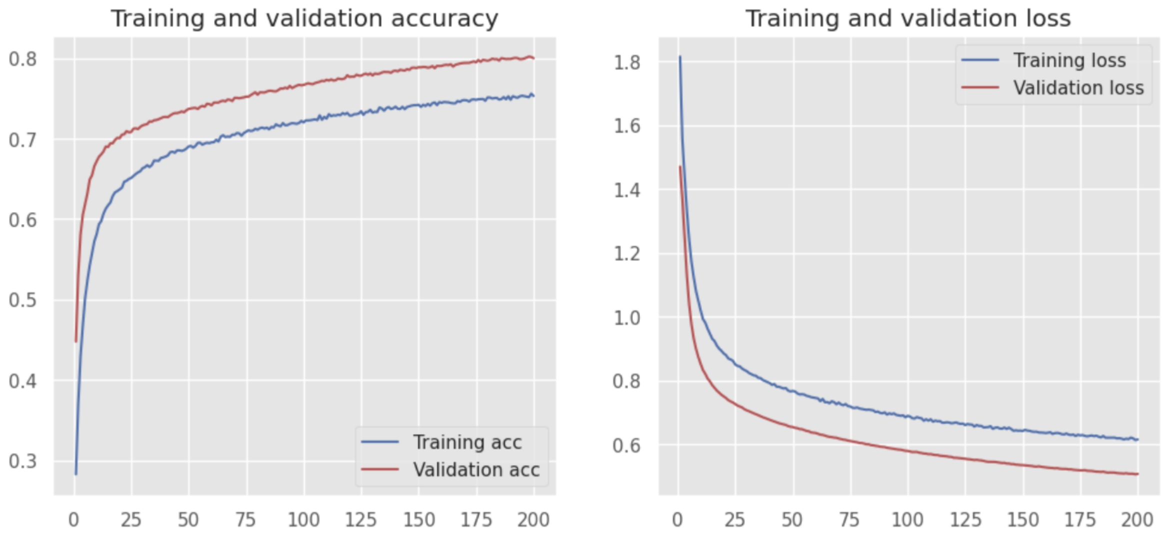
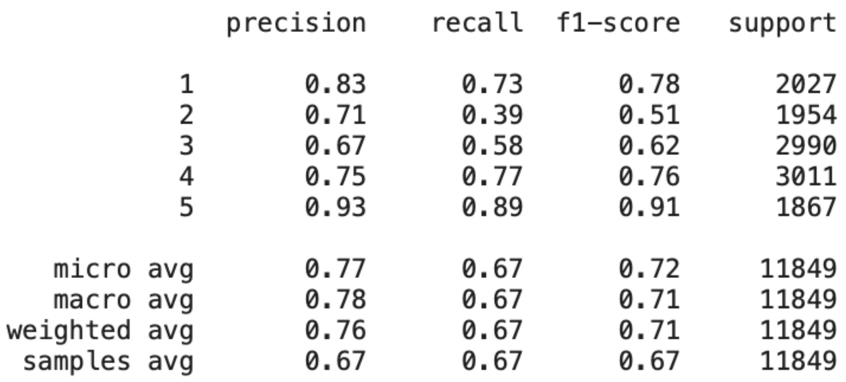


Figure 3: Confusion matrix generated by defined neural network model to predict patient acuity scores



```
               precision    recall  f1-score   support

           1        0.83      0.73      0.78      2027
           2        0.71      0.39      0.51      1954
           3        0.67      0.58      0.62      2990
           4        0.75      0.77      0.76      3011
           5        0.93      0.89      0.91      1867

   micro avg        0.77      0.67      0.72     11849
   macro avg        0.78      0.67      0.71     11849
weighted avg        0.76      0.67      0.71     11849
 samples avg        0.67      0.67      0.67     11849
```

Figure 4: Classification report generated by defined neural network model to predict patient acuity scores

### Discussion of Future Research

For future research and model improvements, efforts to optimize triage classification of middle-tier labels such as 2, 3, 4 may develop the model's effectiveness at answering the research question posed. Focus on refining the chief complaint feature to group certain reasons of admission may provide a more accurate and easy way for a model to distinguish between classes, ultimately providing a more accurate multi-class classification metric between all patient acuity scores. Further feature scaling of patient acuity and re-coding of patient pain scores to include those unable to be assessed and other pain scales may also aid the model in deciphering between triage levels. With re-scaled meaningful features, a model can better generalize the data and provide a more accurate performance of the task. Additionally, ensemble learning may prove to be a beneficial alternative to neural networks when training the model. By combining multiple models, ensemble learning can leverage each model’s strengths and mitigate their weaknesses, resulting in a more accurate and reliable overall prediction.

# References

“Artificial Intelligence-Based Triage. Using AI to Triage Patients in a Healthcare Facility.” Edited by Vlad Medvedovsky, Proxet, 18 June 2021, https://proxet.com/blog/artificial-intelligence-based-triage-using-ai-to-triage-patients-in-a-healthcare-facility/.

Bin Liu, Konstantinos Blekas, and Grigorios Tsoumakas. 2022. Multi-label sampling based on local label imbalance. Pattern Recogn. 122, C (Feb 2022). https://doi.org/10.1016/j.patcog.2021.108294

Johnson, Alistair, et al. “MIMIC-IV-ED.” MIMIC-IV-ED v2.2, 5 Jan. 2023, https://physionet.org/content/mimic-iv-ed/.

Kontio, Elina et al. “Predicting patient acuity from electronic patient records.” Journal of biomedical informatics vol. 51 (2014): 35-40. doi:10.1016/j.jbi.2014.04.001

Pak, Anton et al. “Predicting waiting time to treatment for emergency department patients.” International journal of medical informatics vol. 145 (2021): 104303. doi:10.1016/j.ijmedinf.2020.104303

Sánchez-Salmerón, Rocío et al. “Machine learning methods applied to triage in emergency services: A systematic review.” International emergency nursing vol. 60 (2022): 101109. doi:10.1016/j.ienj.2021.101109

Shafaf, Negin, and Hamed Malek. “Applications of Machine Learning Approaches in Emergency Medicine; a Review Article.” Archives of academic emergency medicine vol. 7,1 34. 3 Jun. 2019

“Timely & Effective Care.” The U.S Centers for Medicare and Medicaid Services, The U.S Centers for Medicare and Medicaid Services, https://data.cms.gov/provider-data/topics/hospitals/timely-effective-care#emergency-department-care.

Wolf, Lisa A et al. “US emergency nurses' perceptions of challenges and facilitators in the management of behavioural health patients in the emergency department: A mixed-methods study.” Australasian emergency nursing journal : AENJ vol. 18,3 (2015): 138-48. doi:10.1016/j.aenj.2015.03.004

Yancey CC, O'Rourke MC. Emergency Department Triage. [Updated 2022 Aug 31]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557583/