The Problem: Enabling Automated Triage for Non-Medical Personnel

People without medical training often lack the prerequisite knowledge or experience to correctly assess when they should seek medical care. Some medical conditions require timely care for good outcomes. These can range from acute emergencies that require immediate care for survival to chronic conditions that should be treated within a few months to avoid severe future consequences. The process for sorting patients into classes based on the priority of care is called triage. Triage is predominantly used by emergency medical personnel to sort patients at mass casualty incidents to maximize the effective use of limited onsite resources and transportation options, but the same principles can be applied to all types of medical conditions. This could be used by medical offices to facilitate outreach from staff to patients or to allow patients to gain insight on how quickly they need to seek care. It is common for patients to hesitate seeking care based on the anticipated expense of seeking emergency room care vs urgent care vs a regular doctors' appointment. Any low cost or self-use option would help fill this gap and be a useful adjunct to the current system. To facilitate better patient triage, a model to quickly classify patients using a series of reported symptoms could be developed and deployed for use in medical apps, doctors' offices, or websites.

The goal is to create a model that sorts conditions into 4 categories:

1) Immediate care needed
    a. Seek medical care as soon as possible,
    b. Activate the local emergency medical system if needed.
2) Seek care soon (~ 2-5 days)
    a. While not immediate, timely care is required to ensure a good clinical outcome.
    b. The disease may progress significantly towards eventual death or disability without the appropriate care.
3) Address at next visit
    a. Condition is chronic and not likely to lead to immediate harm.
    b. The clinical outcome depends on a long-term care strategy with a regular physician, best addressed with medical providers that have or can establish long term relationships with the patient.
4) Need more information for triage
    a. Timeliness and type of care depend on the severity of symptoms that can vary widely.
    b. The next step would be to contact a nurse helpline or a similar service to receive personalized advice.

I will use a dataset collected from a third world location; it contains anonymous information for 4920 patients with 133 symptoms that are used to diagnose 41 diseases. First, I will assign a triage category based on the above criteria for each diagnosis in the dataset. Next, I will use a decision tree model to predict the triage category based on symptom features. I will then fine-tune the model to be as accurate and simple as possible; to eliminate features that would be impossible or difficult for non-medically trained people to assess accurately. I will also attempt to use various techniques to examine the explainability and significance of each feature. This information will then be used to further tune the model for optimal feature selection.

Initial Data Cleaning

I obtained a dataset that contained the medical information for 4920 patients, the dataset includes 133 binary symptom classifiers and 41 medical diagnoses. The dataset had no missing data, except for 1 column (fluid_overload), for which there was a duplicate column (fluid_overload.1). This was fixed by dropping 'fluid_overload' and renaming 'fluid_overload.1' to 'fluid_overload' Additionally, there was an extra unlabeled column where each observation was designated as null that had to be dropped. This was likely an index column that was incorrectly formatted and could not be read correctly into a pandas dataframe.

Oddly, the column for the disease diagnosis was mislabeled as prognosis, which is a medical term that indicates the likely outcome of a condition but is not always a classification per se. This was fixed by renaming the column from "prognosis" to "diagnosis".

Lastly, I generated the test variable (triage category), by assigning each diagnosis to a new triage category indicated by the rules above: 1 = Immediate care needed, 2 = Seek care soon, 3 = Address at next visit, 4 = Need more information for triage.

Exploratory Analysis

Curiously, the dataset had a uniform distribution for the count of each diagnosis (See appendix A). The count for the presence of each symptom was not uniform and ranged from 1932 for 'fatigue' to 102 for 'foul_smell_of_urine'. Some symptom counts were uniform between clusters of symptoms. This tended to occur for symptoms that had lower counts. Unfortunately, this phenomenon is suggestive of multicollinearity. The first decision tree model obtained had a testing accuracy score of 100%, indicated that there was data leakage in the model. Due to the uniformity of certain symptoms, it was highly likely that there was excessive multicollinearity in the data, and that certain symptoms or triage scores were too highly interrelated to be represented as different factors. To assess this, I performed a Cramer's V test across all variables to assess interrelatedness.

Statistical Analysis

The initial Cramer's V test indicated that the model had significant interrelationships between certain symptoms. With multiple pockets of high relationships occurring between certain symptom clusters (see figure 1).

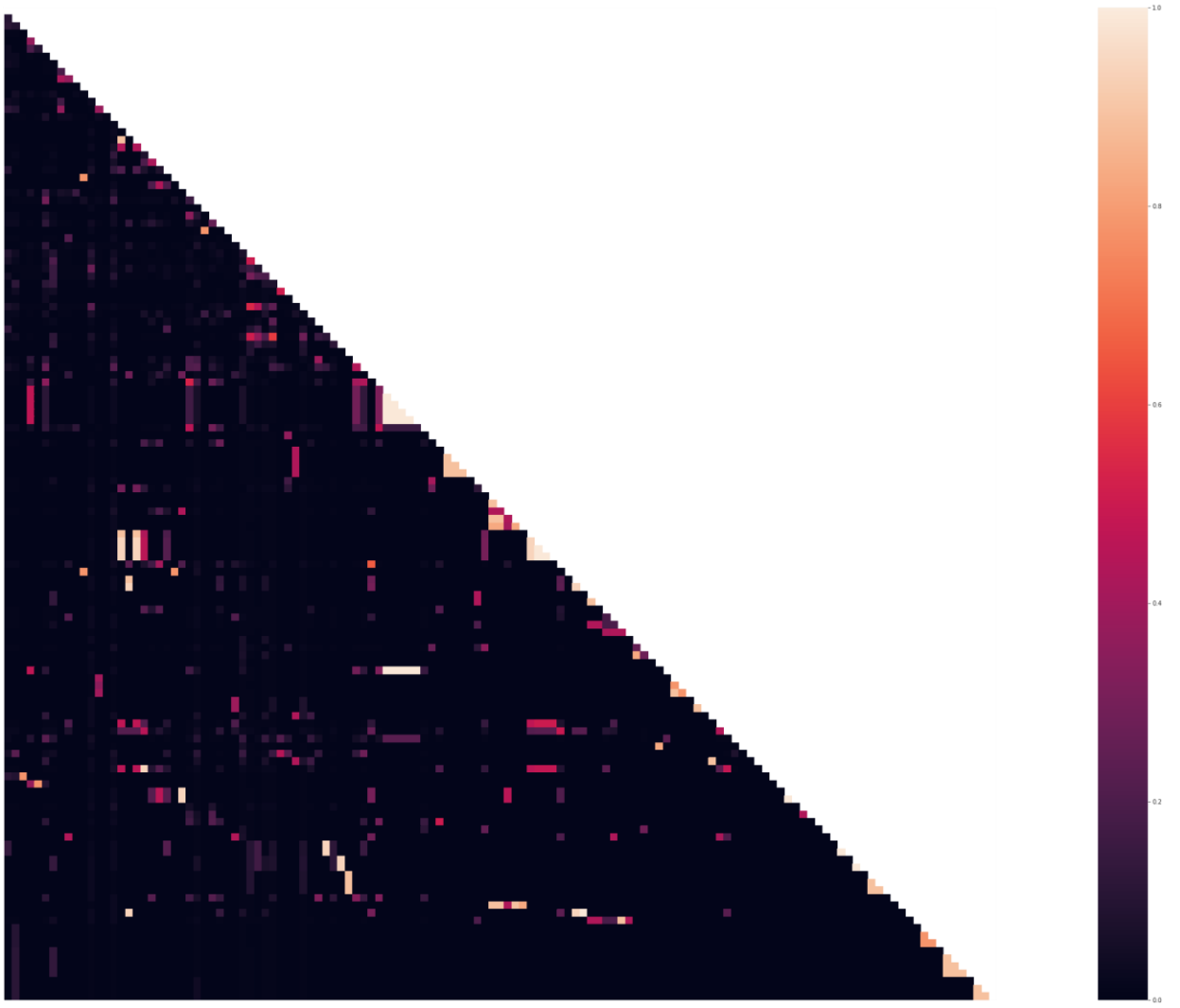


Figure 1: Initial Heat Map Generated Using Cramer's V.

I used the output of the first Cramer's test to remove symptoms that had a strong relationship (> 0.4) with other symptoms using a simple set of logic. I first dropped any symptom that a nonmedically trained person would be unable or highly unlikely to identify correctly (for example 'coma', or 'acidity'). Second I dropped any symptom that was highly related to a set of other symptoms, for example, 'loss_of_smell" would be associated with 'phlegm', 'nasal_discharge', or any other symptom related to a cold or upper respiratory condition. Lastly, I cleaned any binary interrelationship by dropping the symptom that would be more difficult to detect or diagnosis accurately. After applying this set of logical rules, I ran another Cramer's test over the remaining symptoms to confirm there were no longer any symptoms with strong relationships.

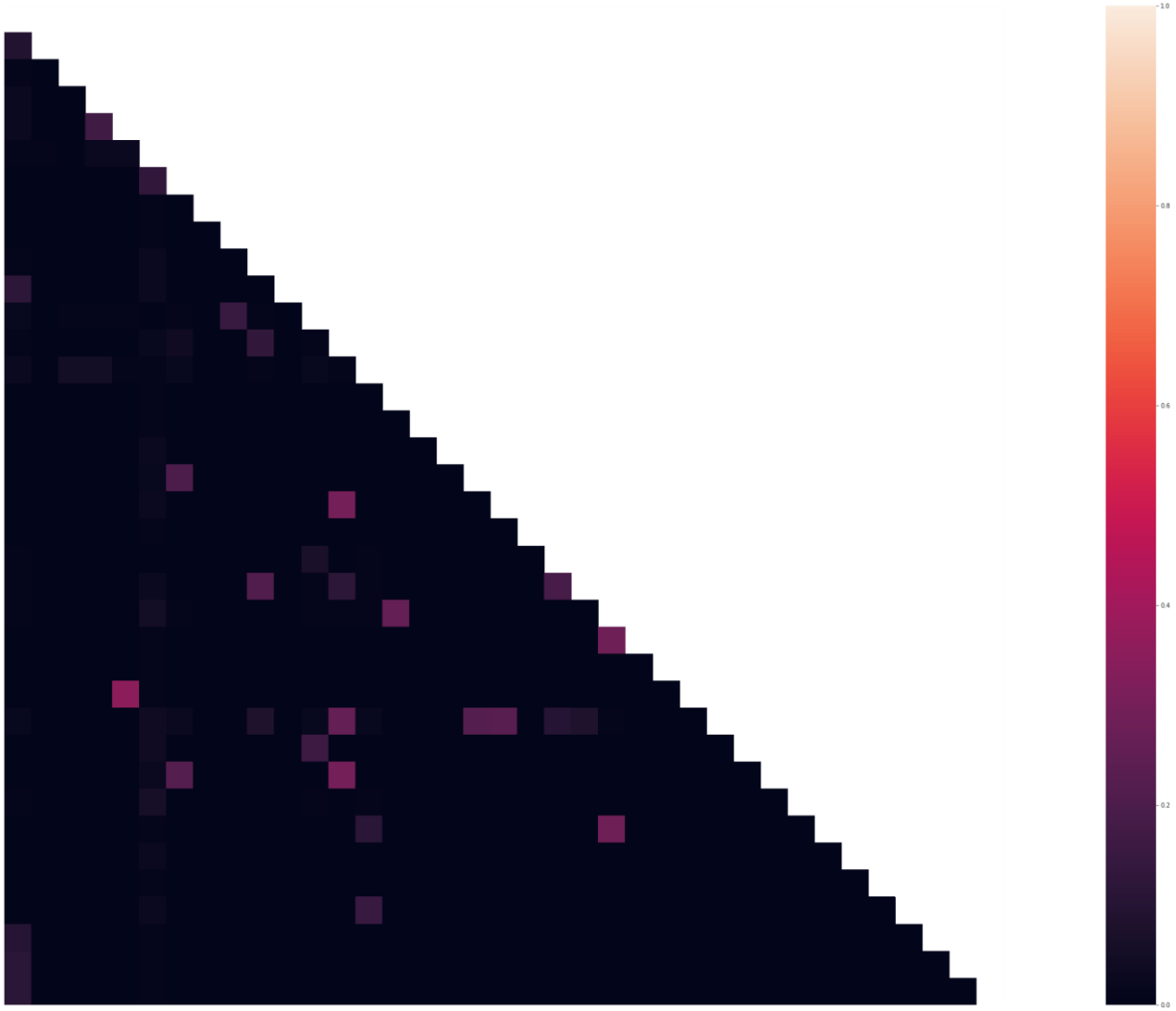


Figure 2: Second Heatmap Generated Using Cramer's Test After Cleaning.

As you can see from the above figure, The second decision trees model using the remaining features no longer had accuracy scores of 100%. Indicating that the data leakage problem had been solved.

Decision Tree Model

I first prepared the data for the decision trees model by splitting the data into predictive features (symptoms) and the target feature (triage classification). I then further split the data into an 80/20 train/test split. The first model was generated without specifying max depth and resolved to 30 levels with 113 nodes. Training accuracy was 90% and testing accuracy was 88%, suggesting slight overfitting of the model, but not enough to conclude that a significant amount of variance occurs. Due to the lack of explainability for a model with so many levels. I generated a new model with a max depth of 3 to see if a less complex model would be accurate enough to be useful.

Table 1: Decision Trees Confusion Matrix for 30 levels

Category	Precision	Recall	F1-score	Support
1: Immediate care needed	0.85	0.78	0.81	268
2: Seek care soon	0.87	0.79	0.83	216
3: Address at next visit	0.99	0.97	0.98	258
4: Need more information for triage	0.83	0.99	0.90	242

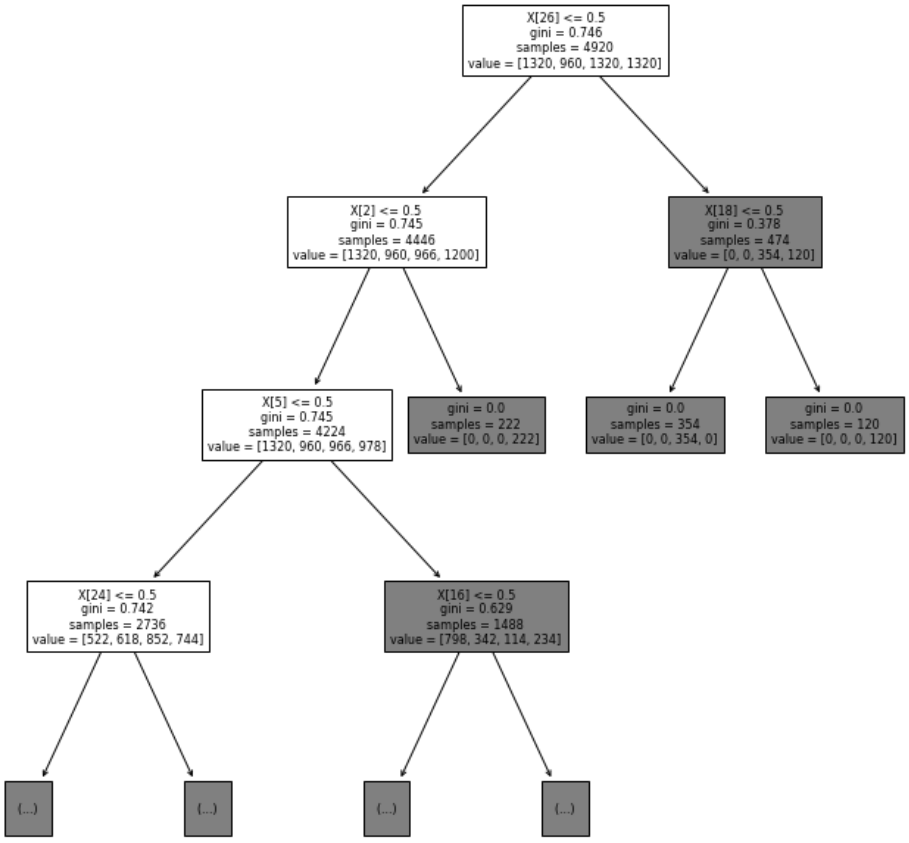


Figure 3: First 3 Layers of the Decision Tree Model

With 3 levels training accuracy was 48% and testing accuracy was 48%. This loss of accuracy is unacceptable. Generating new models with increasing depths was not useful. Even with 10 levels, training accuracy was 76% and testing accuracy was 77%. Since there is no advantage in using a decision trees model that is not explainable. We will evaluate using random forest models instead.

Random Forests Model

I prepared the data for the random forests model by splitting the data into predictive features (symptoms) and the target feature (triage classification). I then further split the data into an 80/20 train/test split. I implemented the random forests model using 10 estimators, and entropy criterion. Training accuracy was 90% and testing accuracy was 88%, suggesting slight overfitting of the model, but not enough to conclude that a significant amount of variance occurs. Dropping further features by least importance did not increase either testing or training accuracy, nor decrease the amount of overfitting. This is the same result achieved by the decision trees model with 30 levels.

Additionally, I generated two models with a max depth of 3 and 10 to compare them with the prior decision trees. For a max depth of 3, the training accuracy was 64% and the testing accuracy was 63% and for a depth of 10 the training accuracy was 86% and the testing accuracy was 85%. This indicates that for this problem a random forests model has superior performance. For this reason, a random forests model without depth restrictions will be used for the final model.

Assessing Random Forests Model Accuracy

The overhaul f1 score for all groups was 0.88. However, there was a significant difference in f1 scores between groups. Category 1 had an f1 score of 0.81, category 2 had an f1 score of 0.83, compared to category 3 with an f1 score of 0.98 and category 4 with an f1 score of 0.90. This indicates the model performs worse for more immediate conditions than conditions that potentially do not require immediate triage.

Table 2: Random Forests Confusion Matrix

Category	Precision	Recall	F1-score	Support
1: Immediate care needed	0.85	0.78	0.81	268
2: Seek care soon	0.87	0.79	0.83	216
3: Address at next visit	0.99	0.97	0.98	258
4: Need more information for triage	0.83	0.99	0.90	242

There were significant differences between precision and recall for each group and between groups. For category 1 the precision was higher (0.85) than the recall (0.78), similar to category 2 that had a precision of 0.87 and a recall of 0.79. This indicates the model is better at finding positives in category 1 and category 2 than correctly classifying them. For category 3 the precision was very high at 0.99 with a slightly lower recall of 0.97. Category 4 had a wide difference between precision (0.83) and recall (0.99), demonstrating a weakness in finding positives, but not correctly classifying them. Overall, the model is better classifying less immediately critical conditions than conditions that may not need immediate care.

Feature Explanation for Final Random Forests Model:

Two methods were used to explain how each feature contributed to the overall model. SHapley Additive exPlanations (SHAP) and recursively dropping each feature and re-implementing the random forest model.

Recursively dropping features indicated that there were 6 features not contributing to the model: indigestion, blurred_and_distorted_vision, weakness_in_limbs, puffy_face_and_eyes, slurred_speech, blood_in_sputum (Appendix C). However, the results of the SHAP disagreed with this finding. Dropping any of these features alone, or in combination, lead to a decrease in model accuracy so they were ultimately maintained.

SHAP analysis indicated that each feature contributed to each classification, but in different amounts (Figure 4). The most common pattern was for one feature to contribute primarily to one class, and then narrowly to the other classes. Another common pattern was for one feature to moderately contribute equally to two classes and then narrowly to the other two. The top 5 most important features using this method were skin_rash, stomach_pain, chest_pain, lethargy, and swollen_legs. The full list of results using this method can be found in Appendix C. Breakdowns of SHAP values for each lass can be found in Appendix D.

Both methods of feature explanation indicate that there are no obvious explanations of how the model classifies the symptoms in triage categories. The model instead uses small parts of each feature holistically to create classifications, with some features being more crucial to certain classifications than others. While this may seem obvious, it is worth noting due to two reasons. First, the drop in accuracy from dropping additional features, indicating that all remaining features are important. Second, the fact that biological systems are highly interlinked due to the evolutionary history of life on earth. Disease and symptoms are presentations of reduced homeostatic potential, with the fallout of one system invariably affecting many other others, potentially leading to signs of stress across the entire organism. Because of this relationships between symptoms that are not detectable in a one on one encounter between a patient and doctor may be detectable at scale when many encounters are analyzed together.

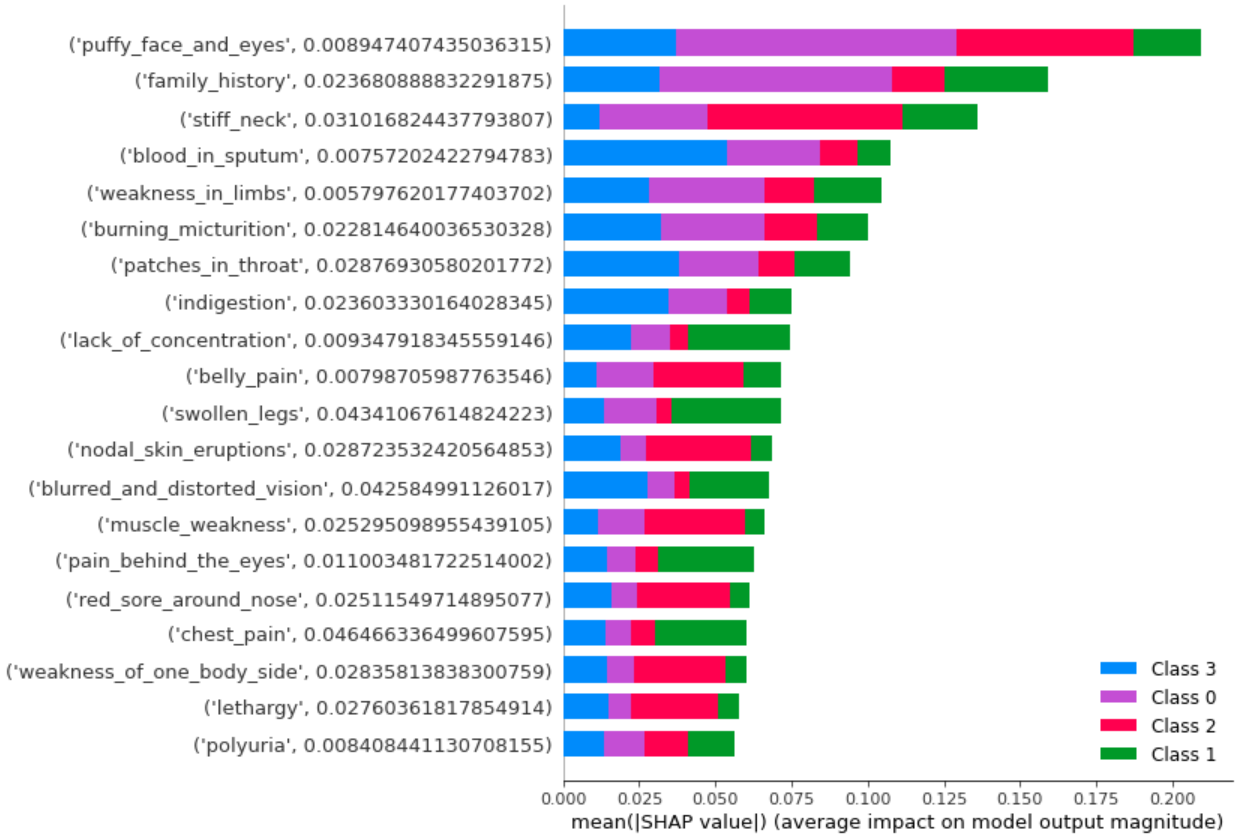


Figure 4: SHAP Impact for Each Category

Next Steps:

The current results of the model are promising enough to move forward with pilot trial of a production system. The main weakness of the model is the uniformity data and uneven precision/recall of the model output. The best way to improve this situation would be to add more data into the model and do parameter tuning until the model has better precision/recall for each class. This could be accomplished by fielding the model in a specific setting under supervision, which would also enable valuable feedback for greater implementation.

Overall, this model is a good first step in demonstrating that this approach to automated triage could be useful. Future systems should be created using more symptoms and diagnoses. As more data is accumulated, the model can be modified to work across a greater range of healthcare settings and roles.

Appendix A: Dataset Diagnoses

Diagnosis	Count
(vertigo) Paroymsal Positional Vertigo	120
AIDS	120
Acne	120
Alcoholic hepatitis	120
Allergy	120
Arthritis	120
Bronchial Asthma	120
Cervical spondylosis	120
Chicken pox	120
Chronic cholestasis	120
Common Cold	120
Dengue	120
Diabetes	120
Dimorphic hemmorhoids(piles)	120
Drug Reaction	120
Fungal infection	120
GERD	120
Gastroenteritis	120
Heart attack	120
Hepatitis B	120
Hepatitis C	120
Hepatitis D	120
Hepatitis E	120
Hypertension	120
Hyperthyroidism	120
Hypoglycemia	120
Hypothyroidism	120
Impetigo	120
Jaundice	120
Malaria	120
Migraine	120
Osteoarthristis	120
Paralysis (brain hemorrhage)	120
Peptic ulcer diseae	120
Pneumonia	120
Psoriasis	120
Tuberculosis	120
Typhoid	120
Urinary tract infection	120
Varicose veins	120
hepatitis A	120

Appendix B: Remaining Symptoms by Frequency

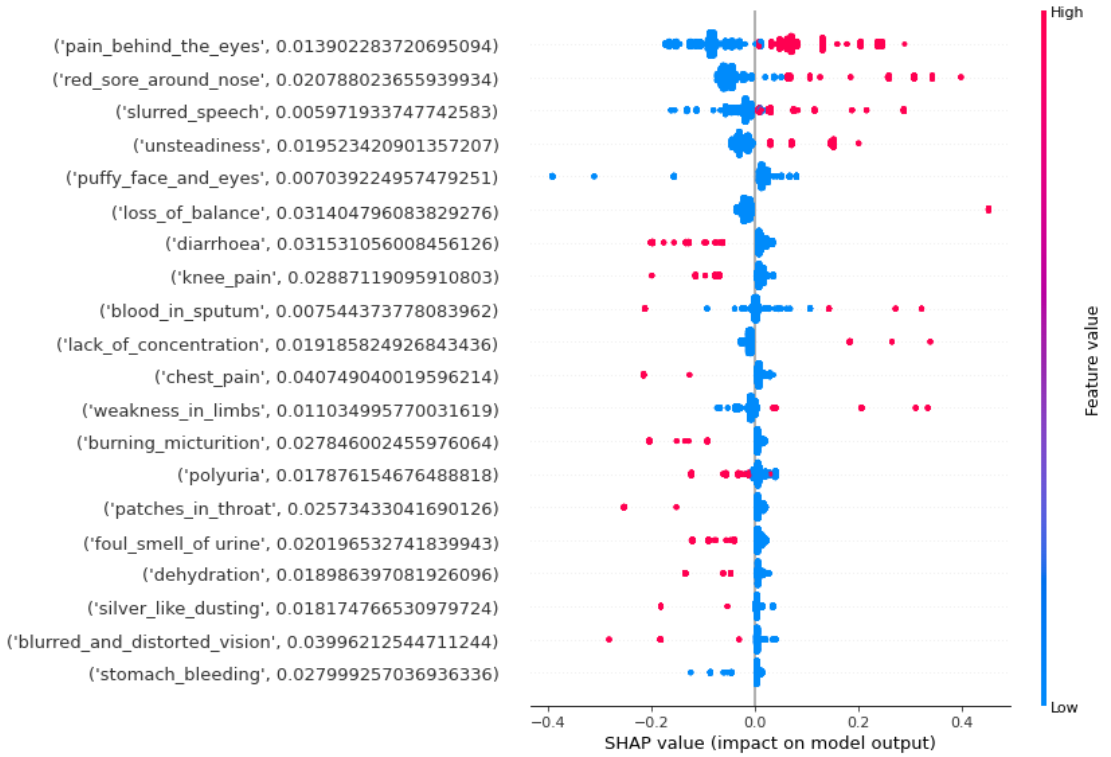
fatigue	**1932**
skin_rash	**786**
chest_pain	**696**
diarrhoea	**564**
irritability	**474**
lethargy	**456**
blurred_and_distorted_vision	**342**
loss_of_balance	**342**
muscle_weakness	**234**
stiff_neck	**228**
family_history	**228**
continuous_sneezing	**222**
stomach_pain	**222**
indigestion	**222**
burning_micturition	**216**
pain_behind_the_eyes	**120**
slurred_speech	**120**
polyuria	**120**
stomach_bleeding	**120**
blood_in_sputum	**120**
pain_during_bowel_movements	**114**
swollen_legs	**114**
puffy_face_and_eyes	**114**
knee_pain	**114**
unsteadiness	**114**
belly_pain	**114**
lack_of_concentration	**114**
history_of_alcohol_consumption	**114**
silver_like_dusting	**114**
red_sore_around_nose	**114**
nodal_skin_eruptions	**108**
patches_in_throat	**108**
dehydration	**108**
weakness_in_limbs	**108**
weakness_of_one_body_side	**108**
pus_filled_pimples	**108**
foul_smell_of urine	**102**

Appendix C: Feature Importances by Recursive Dropping

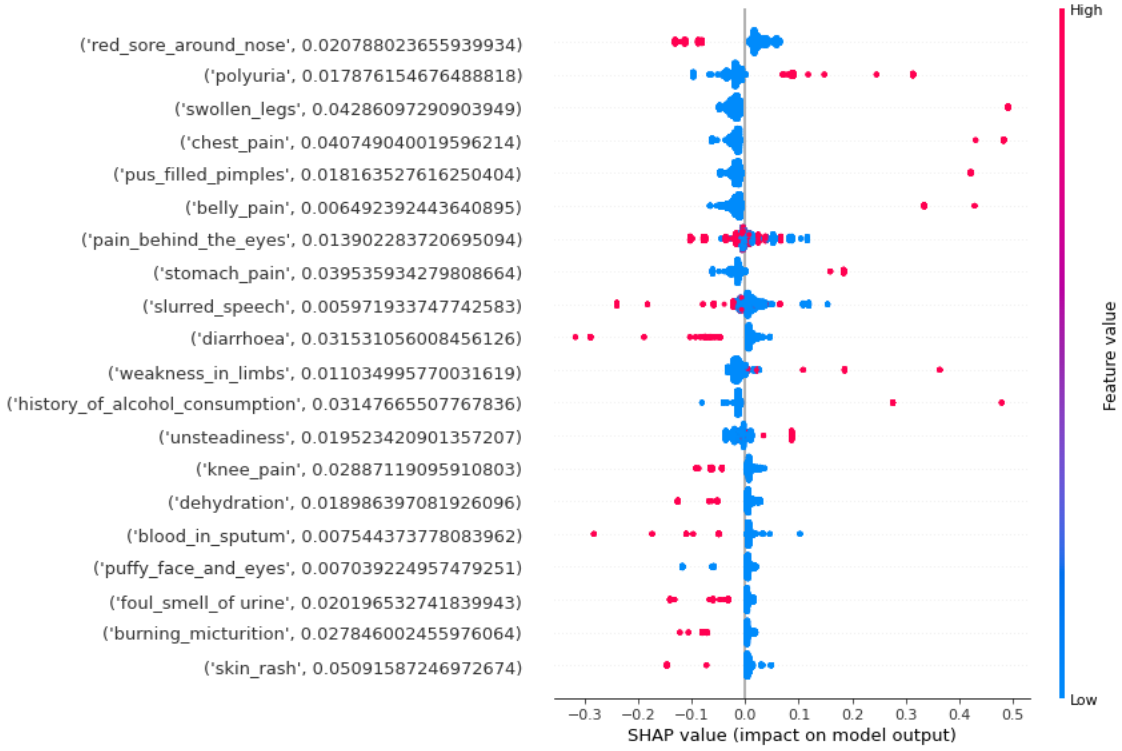
skin_rash	0.028455
stomach_pain	0.025203
chest_pain	0.02439
lethargy	0.023577
swollen_legs	0.023577
knee_pain	0.023577
history_of_alcohol_consumption	0.023577
stomach_bleeding	0.022764
pain_during_bowel_movements	0.021951
diarrhoea	0.020325
patches_in_throat	0.019512
weakness_of_one_body_side	0.019512
continuous_sneezing	0.015447
dehydration	0.013008
family_history	0.009756
burning_micturition	0.007317
fatigue	0.004878
nodal_skin_eruptions	0.003252
red_sore_around_nose	0.002439
foul_smell_of urine	0.001626
irritability	0.001626
pain_behind_the_eyes	0.000813
stiff_neck	0.000813
unsteadiness	0.000813
belly_pain	0.000813
polyuria	0.000813
indigestion	0
blurred_and_distorted_vision	0
puffy_face_and_eyes	0
slurred_speech	0
muscle_weakness	0
lack_of_concentration	0
blood_in_sputum	0
silver_like_dusting	0
weakness_in_limbs	-0.00081
loss_of_balance	-0.00081
pus_filled_pimples	-0.00081

Appendix D: SHAP

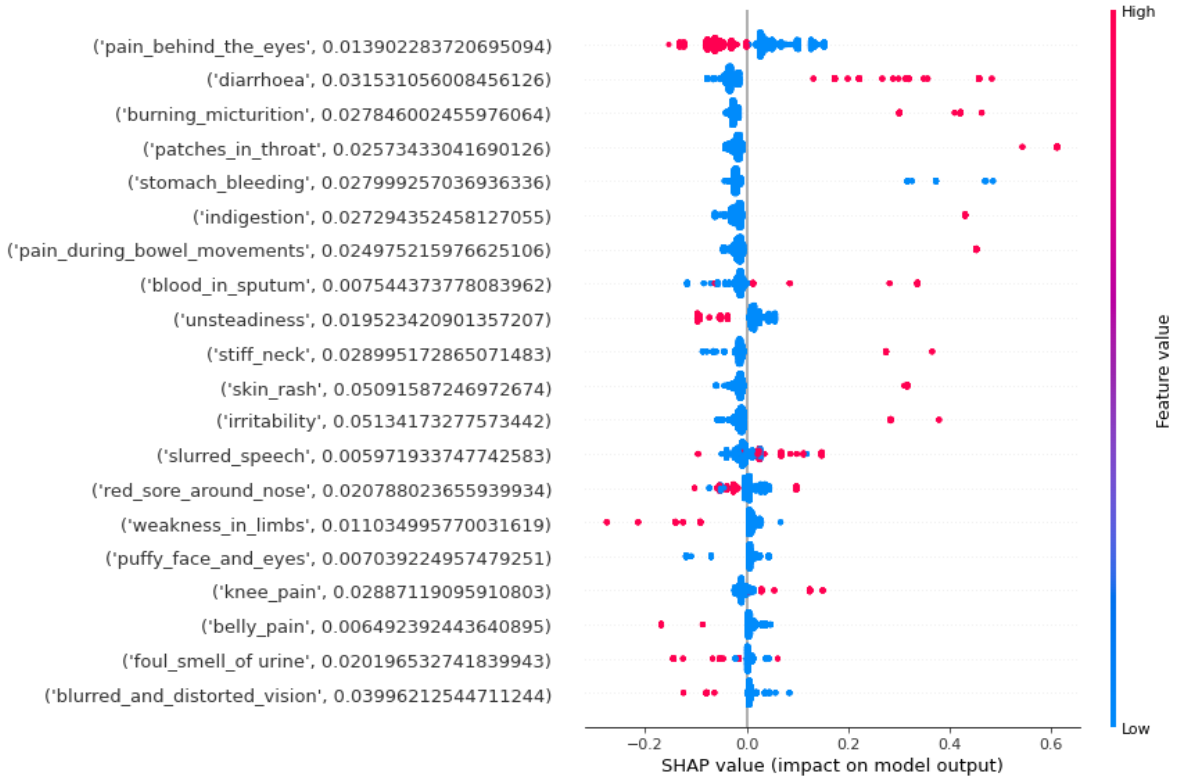
Class 0: immediate care needed



Class 1: Seek care soon



Class 2: Address at next visit



Class 3: Need more information for triage