Symptom Modeling: An N.L.I.C.E. Approach

One of the finidings from the performance of the models was that it was particularly difficult for the models to distinguish between conditions which present similar symptoms and rightfully so. For this project symptoms were modeled as binary values - a value of 1 indicating symptom presentation and 0 denoting an absence of the symptom. This approach to modeling the symptoms was motivated by the nature of the data available in Symcat. However, there is more information that can be obtained about a symptom without having to perform any laboratory or diagnostic tests. This approach to symptom inquiry and the potential for improved performance over the simpler yes-no modeling approach is the focus of this section.

N.L.I.C.E: An Introduction

Once a symptom is known to be present, there are additional information which can be obtained about the symptom. N.L.I.C.E is an abbreviation for these additional data.

Nature

It is possible to also obtain the nature of the presented symptom. As an example, consider cough which is a common symptom associated with respiratory infections. A 1 or 0 approach to modeling this symptom ignores the different ways cough might present. Some patients might experience a dry cough, others might have cough which produces mucus, etc. These differences define the nature of the cough and provide information which might make it easier to distinguish between conditions which have similar symptoms.

Location

The location of a symptom can also be a discriminating factor when making distinctions between conditions. As an illustration consider a patient who complains of abdominal pain. In medical anatomy, the abdomen itself can be divided into four quadrants - upper and lower right quadrants as well as upper and lower left quadrants. There is also a division into 9 regions - right and left hypochondriac regions, right and left lumbar region, right and left iliac regions, as well as the epigastric, umbilical and hypogastric regions. Additional information on which regions/quadrants exactly in the abdomen are affected by this pain might be helpful in making better distinctions regarding the causal condition.

Intensity

Symptom intensity i.e how severe the symptom presentation is, can also be a pointer to the causal condition. Continuing with the abdominal pain example, a patient suffering from appendicitis is more likely to experience severe to moderate abdominal pain than a patient who suffers from irritable bowel syndrome.

Chronology

To aid an explanation of this term, chronology can be split into sub terms: frequency - how often the symptom occurs, duration - how long the symptom persists and onset - when the patient started experiencing the symptom. These could also be additional discriminative factors when making a differential diagnosis.

Excitation

Excitation refers to activities or scenarios that the patient performs or is exposed to which triggers or worsens the symptoms. As an example a patient might complain of leg pain only while running. Such information, if captured might also make for a more accurate distinction between conditions.

Data Collection and Generation

As mentioned earlier, the data provided in Symcat was not immediately suited to an N.L.I.C.E. modeling approach. Some symptoms present in Symcat did capture some components of this approach e.g there were conditions which presented with *burning-abdominal* pain and others with *sharp-abdominal* pain (both distinctions made on the nature of the abdominal pain). However, this was not available for all the symptoms which could support the N.L.I.C.E approach. Hence, as a proof of concept, data for a small set of conditions was gathered from medical literature. This data gathering effort was carried out by medical experts from our industry collaborators at Medvice. For the data collection, the conditions were restricted to four categories based on the part of the body which the conditions affect. The conditions are shown in table [table: nlice-conditions]

Table 1: Selected Conditions for NLICE Analysis

Condition Group	Condition Name
	Acute Pancreatis
	Appendicitis
	Functional Constipation
	Irritable Bowel Syndrome

Condition Group	Condition Name
	Migrane
	Tension Type Headache
	Subarachnoideal Bleeding
	Covid-19
	Seasonal Influenza
	Pneumothorax
	Hernia Nuclei Pulposi
	Lumbago / Muscle strain
	Cauda Equina

For each of this conditions, data regarding typical symptoms, associated expression probability, as well as NLICE characteristics were extracted manually from medical literature. Using the generation method described in chapter [chapter: 3] two data sets were generated based on the collected information: a basic-dataset where the symptoms were modeled with the binary model approach and another with the N.L.I.C.E. approach.
It should be stated that symptoms do not have to support all NLICE characteristics. As an example, fever can be characterized by its intensity but it would not be logical to associate a location to fever.

Feature Encoding

The 1 or 0 encoding for symptom presence/absence was still maintained. A one hot encoding scheme was applied to the *Nature*, *Location*, *Intensity* and *Excitation* characteristics. A *not-available* category was used to encode cases where the symptom either did not support an NLICE characteristic or data was not available for that particular symptom. Chronology was split into three features: frequency - for which a one hot encoding scheme was used, duration and onset which were both expressed in seconds.

Model Training and Results

Following the same approach as outlined in chapter [chapter: 4], a Naive Bayes and Random Forest model were trained on both the basic-dataset and the NLICE dataset. Table [table: nlice-baseline-compare] shows the results obtained by both models on these datasets.

Table 2: Selected Conditions for Confusion-Comparison

	Accuracy	Precision	Top 5 Accuracy	Accuracy	Precision	Top 5 Accuracy
Basic	0.911	0.914	1.000	0.915	0.916	1.000

Table 2: Selected Conditions for Confusion-Comparison

NLICE	0.958	0.964	1.000	0.955	0.956	0.999

Comparing the results we see an improvement in accuracy and precision compared to the basic dataset. The unusually high accuracy levels in both cases can be attributed to the small size of the condition and symptom space. An intuitive explanation is that the models simply have less sources of confusion and are better able to make distinctions between conditions.
The increased performance however suggests that there is a potential benefit to the NLICE approach and if extended to as many conditions and symptoms as contained in Symcat might also yield better results.