Healthcare Symptom Classification Using Classical and Neural NLP Models

Islah Haoues - 1800272

Baker Huseyin - 1901345

1. Introduction

Predicting diseases from patient-reported symptoms is a challenging task due to symptom overlap across conditions and limited discriminative signal in short textual descriptions. In this project, we study a healthcare symptoms dataset in which each record consists of a small set of symptoms associated with a diagnosed disease. The original task formulation involves predicting one of 30 disease classes from symptom text.

Preliminary analysis revealed that the original 30-class problem is not learnable due to extreme overlap in symptom distributions across diseases. To address this limitation, we apply data-driven disease grouping to reduce label entropy and enable meaningful supervised learning. We evaluate four models of increasing complexity: a classical machine learning baseline using TF-IDF features, a feed-forward neural network using multi-hot symptom vectors, a simple recurrent neural network, and a long short-term memory network.

2. Dataset Description and Preprocessing

The dataset contains approximately 25,000 samples, each consisting of patient metadata and a short comma-separated list of symptoms. Across the dataset, there are 28 unique symptoms and 30 unique diseases. Each sample contains between three and seven symptoms.

Symptoms are normalized by lowercasing and trimming whitespace. Three representations are derived from the symptom lists: textual representations for TF-IDF modeling, multi-hot binary vectors for feed-forward neural networks, and integer-encoded sequences for recurrent models. These representations are used consistently across all experiments.

3. Disease Grouping via Unsupervised Clustering

Exploratory analysis showed that individual symptom frequencies are nearly identical across all 30 diseases, resulting in negligible mutual information between symptoms and disease labels. As a result, all models trained on the original label space converged to near-random performance.

To reduce label entropy while preserving statistical structure, a disease-level symptom frequency matrix was constructed, where each disease is represented by a normalized distribution over symptoms. K-means clustering was applied to this matrix with $k = 6$. This initial clustering produced two singleton clusters corresponding to outlier diseases. These singleton clusters were reassigned to their nearest non-singleton cluster based on Euclidean distance to cluster centroids.

After this reassignment, the final label space consisted of four disease groups. These four groups were used as the target variable for all subsequent models. The resulting class distribution is reasonably balanced, and the random baseline accuracy for this setting is 25%.

4. Models

Four models were implemented and evaluated using identical train, validation, and test splits.

1. The first model is a classical machine learning baseline using TF-IDF features extracted from symptom text and a linear support vector machine classifier.

2. The second model is a feed-forward neural network trained on 28-dimensional multi-hot symptom vectors. The network consists of two hidden layers with ReLU activations and dropout for regularization.

3. The third model is a simple recurrent neural network trained on integer-encoded symptom sequences using an embedding layer followed by a recurrent layer.

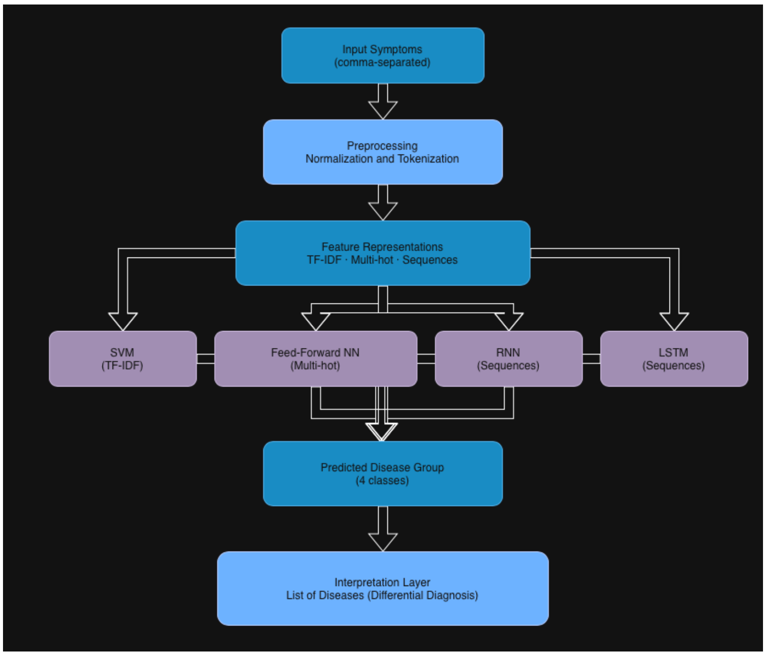
4. The fourth model replaces the recurrent layer with an LSTM layer to capture longer-range dependencies within symptom sequences.

All neural models are trained using categorical cross-entropy loss and the Adam optimizer.

5. Experimental Setup and Architecture

The dataset is split into training, validation, and test sets using stratified sampling to preserve class proportions. Model selection is performed based on validation accuracy, and final results are reported on the held-out test set.

The diagram on the right shows the architecture of the prediction pipeline.



6. Results

All models achieved performance above the random baseline. The SVM with TF-IDF features achieved a test accuracy of 28.7 percent. Neural models performed slightly better, with the feed-forward neural network achieving 30.1 percent accuracy. Sequence-based models further improved performance, with the simple RNN achieving 30.5 percent and the LSTM achieving the highest test accuracy of approximately 30.6 percent.

While differences between models are modest, the consistent improvement from classical to neural and sequence-based models indicates that limited but meaningful structure exists in the grouped label space.

Model	Input Representation	Validation Accuracy	Test Accuracy
SVM (TF-IDF)	Symptom text	0.2851	0.2872
FFNN	Multi-hot vectors	0.2923	0.3008
RNN	Embedded sequences	0.2931	0.3048
LSTM	Embedded sequences	0.2947	0.3051

7. Dataset Limitations and Critical Analysis

A key finding of this project is that the dataset structure fundamentally constrains achievable performance, regardless of model architecture. After normalizing symptom lists through sorting and deduplication, more than 21,000 of approximately 25,000 rows contain unique symptom combinations. This extreme sparsity prevents models from encountering repeated symptom patterns during training, making generalization difficult.

Additionally, symptom frequency analysis reveals that each symptom appears with nearly identical probability across diseases. This lack of discriminative signal explains why all models initially converged to random performance under the original 30-class formulation. Even after extensive experimentation with alternative preprocessing strategies, embeddings, and model architectures, results remained unchanged until the label space was restructured.

Grouping diseases into four data-driven categories introduces limited but measurable separability, raising accuracy to approximately 30 percent. However, all models converge to similar performance levels, indicating that model complexity is not the primary limiting factor. Instead, performance is constrained by the inherent structure of the dataset.

Across all models, sequence-based neural architectures marginally outperform classical baselines, though gains remain limited. During demonstration, predictions are presented as disease groups along with the list of diseases contained within each group, providing an interpretable and clinically realistic differential diagnosis rather than an artificially precise single-disease prediction.