

Class-Conditional Conformal Prediction on Multilabel Classification of EHR Data

Daniel Tcheurekdjian¹, Andy Liu², Joshua Klasmeier¹, and Torey Hilbert²

1. Aleph Innovations and 2. The Ohio State University

Introduction

Abstract

Deep Learning (DL) methodologies have been shown to be effective in improving natural language processing tasks across disciplines. Recent developments in the field have advanced DL efficacy at the sentence level, including natural language inference, and at the token level, including named entity recognition and question answering.

These advances have failed to materialize in disciplines with complex domain-relevant information structures, such as medicine, as current architectures struggle to classify highly complex outputs with strict guarantees of no information corruption.

We present a neural architecture composed of a hierarchical series of networks which allow us to restrict the search space for a given classification task, allowing for easier and faster convergences during optimization, and higher model accuracies across validation sets.

We further use conformal prediction to make provable guarantees about the error rates of the model.

Conformal Prediction Overview

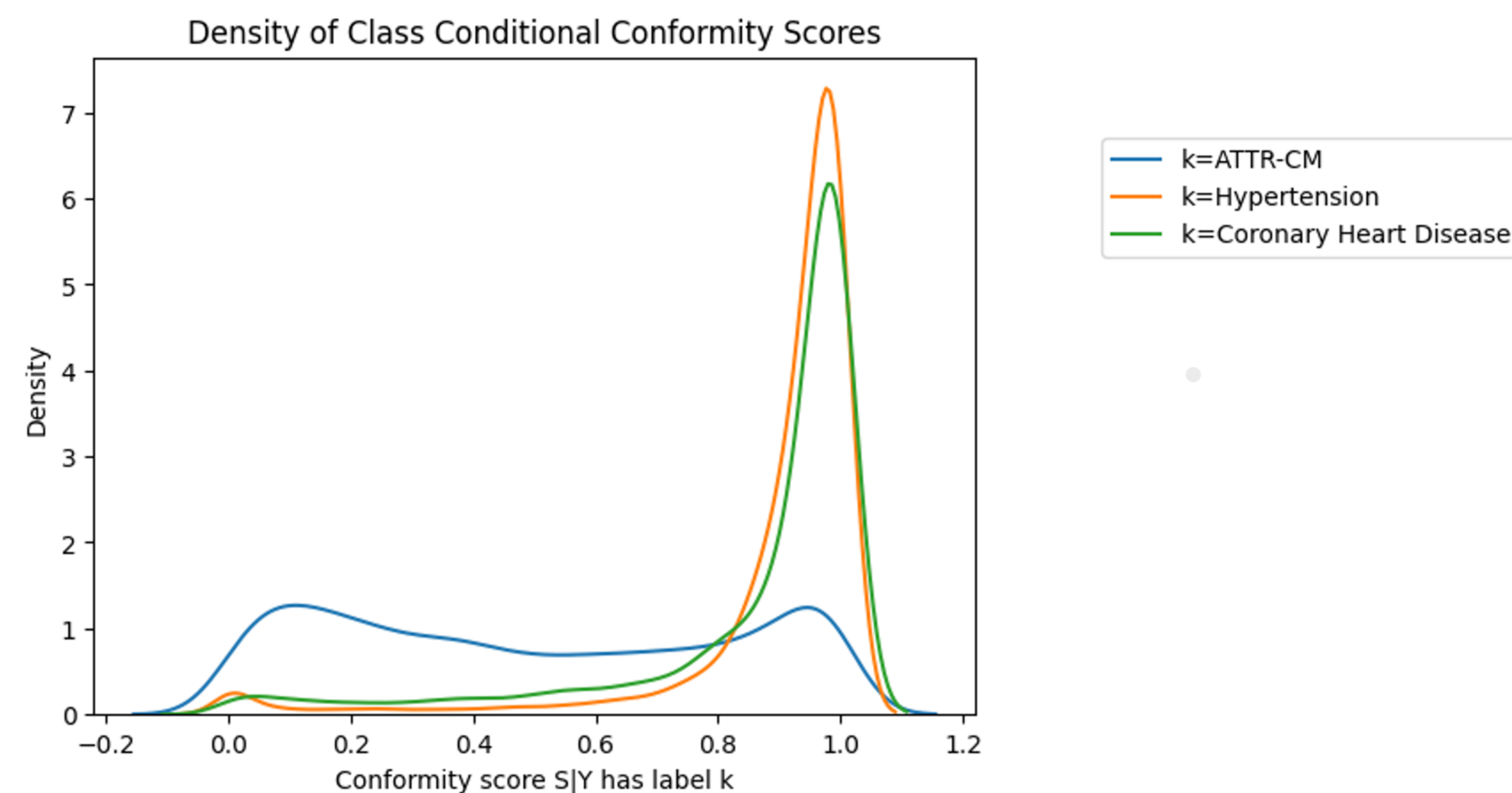
Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)(X_{n+1}, Y_{n+1}) \stackrel{iid}{\sim} \mathbb{P}_{x,y}$

Suppose we have a model f that we want to calibrate.

Let $S_i = 1 - (f(x_i))_{Y_i}$, then for any k , assuming for simplicity's sake, all S are distinct, then we have,

$$\begin{aligned} \frac{k}{n+1} &= \mathbb{P}(S_{n+1} \geq k^{th} \text{ largest of } \{S_1, S_2, \dots, S_n\}) \\ &= \mathbb{P}((f(X_{n+1}))_{Y_{n+1}} \geq 1 - k^{th} \text{ largest of } \{S_1, S_2, \dots, S_n\}) \\ &= \mathbb{P}(Y_{n+1} \in \{j : f(X_{n+1})_j \geq 1 - k^{th} \text{ largest of } \{S_1, S_2, \dots, S_n\}\}) \\ &= \mathbb{P}(Y_{n+1} \notin \{j : f(X_{n+1})_j < 1 - k^{th} \text{ largest of } \{S_1, S_2, \dots, S_n\}\}) \\ &= \mathbb{P}(Y_{n+1} \notin \hat{C}(X_{n+1})), \end{aligned}$$

where $\hat{C}(x_{n+1}) = \{j : f(X_{n+1})_j > 1 - k^{th} \text{ largest of } \{S_1, S_2, \dots, S_n\}\}$.



Class-conditional Conformal Prediction

Class-conditional Conformal Prediction

Conformal prediction has an inherent flaw. Consider the following classification task: we have a case where our error rate is 5%, 99% of our true labels are negative and 1% of our true labels are positive. Despite having a nominal error rate of 5%, this model has an inaccuracy rate of 100% when the true label is positive.

To avoid this problem, we calibrate our prediction set using a set of purely positive and purely negative cases, following roughly the same process as in general conformal prediction to guarantee $1 - \alpha \leq \Pr(Y_{test} \in C(X_{test}) | Y_{test} = y) \leq 1 - \alpha + \frac{1}{n+1}$ where $C(x_{test}) \subset (1, \dots, K)$.

Choose a label j . Let $S_1^j, S_2^j, \dots, S_{n_k}^j$ be the n_k scores S_i such that $k \in Y_i$. In other words, we are considering $\{(X_i, Y_i) : Y_i \text{ contains the label } k\}$.

Then we have:

$$\begin{aligned} \frac{k}{n+1} &= \mathbb{P}\left(S_{n+1} \geq k^{th} \text{ largest of } \{(X_i, Y_i) : Y_i \text{ contains the label } k\} \right. \\ &\quad \left. | Y_{n+1} \text{ contains the label } k\right) \\ &= \mathbb{P}\left(Y_{n+1} \notin \{j : f(X_{n+1})_j < 1 - k^{th} \text{ largest of } \right. \\ &\quad \left. \{(X_i, Y_i) : Y_i \text{ contains the label } k\} | Y_{n+1} \text{ contains the label } k\}\right) \\ &= \mathbb{P}(Y_{n+1} \notin \hat{C}(X_{n+1}) | Y_{n+1} \text{ contains the label } k), \end{aligned}$$

where $\hat{C}(x_{n+1}) = \{j : f(X_{n+1})_j > 1 - k^{th} \text{ largest of } \{S_1^j, S_2^j, \dots, S_{n_k}^j\}\}$.

Data Collection

The Massachusetts Institute of Technology Laboratory for Computational Physiology, in collaboration with the Beth Israel Deaconess Medical Center, prepared a collection of deidentified free-text clinical notes for patients included in the MIMIC-IV clinical database: MIMIC-IV-Note.

MIMIC-IV-Note contains 331,794 deidentified discharge summaries from 145,915 patients admitted to the hospital and emergency department at the Beth Israel Deaconess Medical Center in Boston, MA, USA. All notes have had protected health information removed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision.

Each note contains the full record of a patient's course-of-care at the Beth Israel Deaconess Medical Center, and the total dataset includes over 200,000 unique diseases, resulting in over 300,000 unique disease states. This implies that each diagnostic classification task for an admitted patient is unique.

The data was further processed to correct misspellings, close unnecessary spaces, expand abbreviations, and fill in acronyms.

Neural Architecture and Outputs

Architecture

Our model consists of a cascade of neural networks, each level classifying EHRs with increasing granularity. We employ a hybrid design, leveraging the strengths of both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

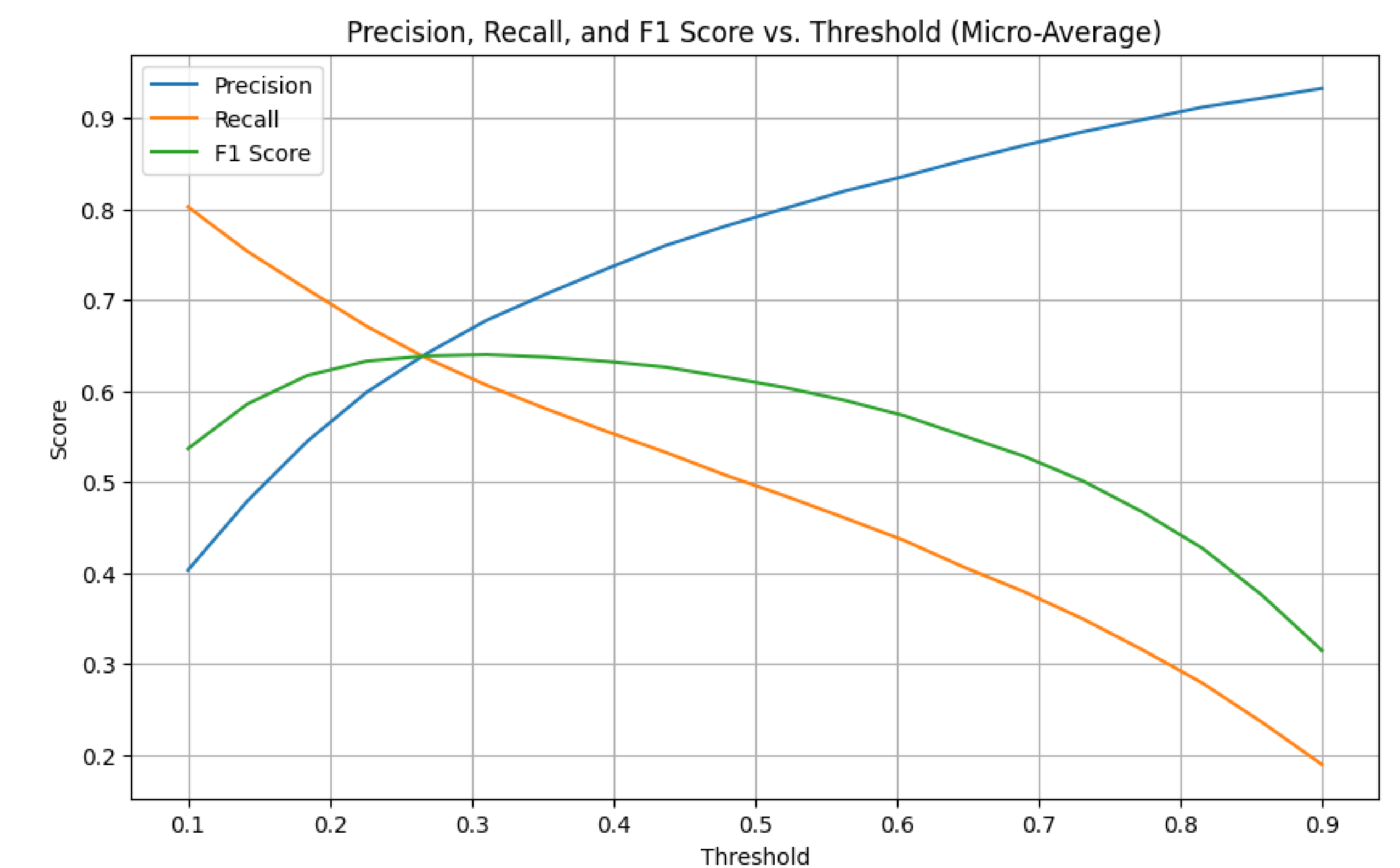
We first use a CNN to extract high-level features from raw EHR data. The features are fed into subsequent networks in the hierarchy, which use RNNs with LSTM units to refine the classification further. This neural architecture intentionally mimics the way human clinicians analyze medical data.

By classifying into progressively finer categories, we dramatically reduce the computational search space that the model needs to explore at each level. This leads to faster training convergence and improved generalization.

Results

The model's performance was assessed at three barriers: classification of board, classification of subspecialty, and classification of disease.

Disease classification metrics are plotted below. In general, recall remained low across all 3 barriers for our chosen threshold (0.5), precision dropped as predictions became more granular (Means: 94, 87 80), and convergence speeds increased (Epochs: 100, 80, 17). Furthermore, the analysis of class conditional conformity scores showed a high density around 1 for most classes (83%).



Reference: Angelopoulos, A. N., & Bates, S. (2022, December 7). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. arXiv.org. <https://arxiv.org/abs/2107.07511>
Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023, January 6). *Mimic-IV-note: Deidentified Free-text clinical notes*.