Bayesian Priors From Large Language Models Make Clinical Prediction Models More Interpretable

Avni Kothari MS [1], Daniel J. Bennett MD[1], Seth Goldman MD[1], Elizabeth Connelly MPH[1], James D. Marks MD, PhD[1], Lucas S. Zier MS, MD[1], MS, Jean Feng MS, PhD[1]
[1]University of California at San Francisco

Introduction Electronic health record (EHR) data provides the opportunity to create highly impactful clinical risk prediction models. However, an issue with deploying a machine learning (ML) model trained on thousands of features extracted from the EHR is that the model often learns to rely on spurious features because many non-causal features can be highly correlated. Despite achieving high performance in internal validation studies, these models often underperform in external validation studies and can be challenging to adopt. The importance of having a feature set that is clinically meaningful is crucial for a clinician to understand and trust risk predictions from a ML model. To ensure clinical model interpretability, the predominant approach is to manually curate a set of features. However, the time and effort needed for clinicians to sift through thousands of features is substantial; therefore, clinicians typically enumerate a small set of features which may not be highly predictive in aggregate.

To address the issue of feature interpretability, we leverage the vast knowledge of large language models (LLMs). Our approach efficiently identifies which features extracted from EHR are likely to be predictive and linked to the outcome of interest, mimicking the behavior of an experienced clinician. Outputs from the LLM are then used to regularize the ML model to be clinically interpretable. While previous work has used LLMs to select features [1], this approach can degrade model performance since the LLM is error-prone. Instead, our framework seeks to incorporate LLM-generated feature rankings in a principled manner through Bayesian inference: the LLM-generated prior is used as feature prior weights for training a Bayesian Additive Regression Trees (BART) model. BART is a Bayesian formulation of gradient boosted trees; it learns a posterior distribution over the space of tree ensembles given a prior over tree structures [2]. It can reconcile imperfect clinical knowledge embedded in a LLM with empirical data to train models that depend on features that are clinically relevant *and* highly predictive.

We demonstrate how the LLM+BART framework can be used to revise a 30-day all-cause readmission model for heart failure patients developed by the Zuckerberg San Francisco General Hospital (ZSFG) predictive analytics team. The team's initial model was trained on 3451 features without clinical curation, which resulted in an uninterpretable model with features likely serving as proxies for more clinically meaningful concepts. Using LLM+BART, we show that the revised model is more interpretable while achieving the same performance.

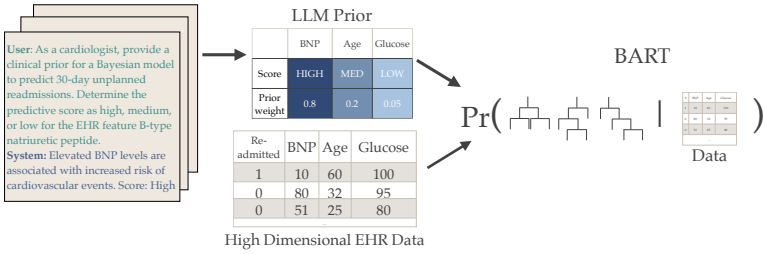

Figure 1: Demonstration of the LLM+BART framework. For features in the EHR the LLM is prompted to provide a feature score to be used as prior in training the BART model.

Methods Consider a classification task with training data $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{n}$, where each observation is associated with a $\boldsymbol{d}$-dimensional feature vector $\boldsymbol{x}^{(i)}$ and binary label $\boldsymbol{y}^{(i)}$ and we have metadata for all $\boldsymbol{d}$ features. The metadata may be text defining the feature or simply the feature name itself. To make our framework more concrete, we will refer back to the ZSFG readmission classification task as an example, where the only metadata available was the column name or attribute name in the Epic EHR system.

Obtaining feature priors**:** For all features we query a LLM for a score ranking how predictive and clinically relevant each feature is, where the possible scores are "high" (3), "medium" (2), and "low" (1). To provide context to the LLM, each prompt begins with a few examples for features in each category along with some short reasoning, a technique known as in-context learning. For example, in the readmission task, we categorize "b-type natriuretic peptide (BNP)" as high because of previous literature linking elevated levels at discharge with an increased 30-day readmission risk [3], echocardiograms measurements such as LVOT VTI as medium risk because it is a surrogate for cardiac output yet has not been directly linked to 30 day readmission risk , and "serum prostate specific antigen" as low because it is unrelated to readmissions. Note that these examples serve as reference points rather than ground truths for the LLM since the LLM-generated prior can be corrected by the BART model. After, the LLM is presented with metadata for

the feature of interest and asked to employ clinical reasoning to determine the feature's score, a technique known as chain-of-thought prompting.

Bayesian inference for model fitting: BART posits the data to be distributed per the sum of $\boldsymbol{K}$ decision trees and embeds this model within a Bayesian framework to support uncertainty quantification and provide a principled approach to regularization through prior specification. More specifically, the data is assumed to be distributed per $\mathbf{Pr}(\boldsymbol{Y}=\mathbf{1}|\boldsymbol{X}) = \frac{\mathbf{1}}{\mathbf{1}+\mathbf{exp}(-\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \boldsymbol{f_k}(\boldsymbol{x}))}$ for decision trees $\boldsymbol{f_1}, \cdots, \boldsymbol{f_K}$ and the tree prior for splitting on the $\boldsymbol{j^{th}}$ feature follows a Dirichlet distribution with feature weights $\boldsymbol{w_k}$ i.e. $\boldsymbol{P_{split}}(\boldsymbol{j}) \propto \prod_{\mathbf{k}=\mathbf{1}}^{\mathbf{d}} \mathbf{w_k}^{\mathbb{1}\{\boldsymbol{k}=\boldsymbol{j}\}}$. Rather than using the default split prior that is uniform across all features, we encourage the trees to split on features that are believed to be clinically relevant by defining the feature weights to be proportional to their score from the LLM. BART is then fit using Markov Chain Monte Carlo (MCMC), which combines the empirical data and the feature priors to obtain a posterior distribution over tree ensembles as showing in Figure 1. By conducting Bayesian inference, BART regularizes the model towards more clinically interpretable models and determines if erroneous priors from the LLM need to be corrected based on the data. This powerful combination allows us to create high-performing models with an interpretable set of features.

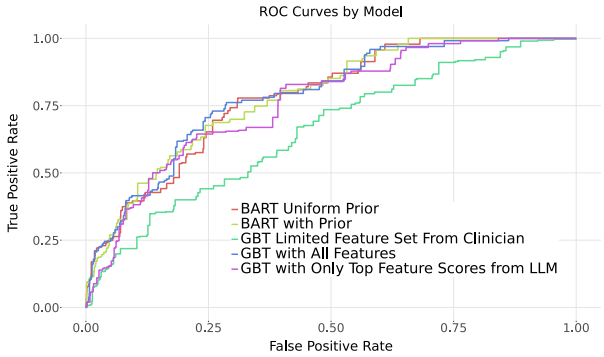

Figure 2: Performance of ML models with uniform priors versus LLM+BART. LLM+BART matches the best performing models while being more interpretable.

Results We apply our LLM+BART framework to revise an all cause 30-day readmission model for heart failure patients at ZSFG. The original algorithm trained on 3210 encounters was a gradient boosted classifier (GBT) that achieved a test AUC of .78. However, the top features (Figure 3) were found to be uninterpretable and non-intuitive.

In our framework we utilized Llama-13B, a general-purpose open-source LLM trained on publicly available online data sources including Wikipedia. We compare the LLM+BART model by fitting (a) a GBT model with all EHR features, (b) a GBT on only the top features (scores greater than 2) [1], (c) BART with uniform feature priors, and (d) GBT only using a limited feature set curated by clinicians on the ZSFG predictive analytics team.

The top performing methods were the GBT classifier with all EHR features, the BART model with uniform feature priors, and the LLM+BART model, which all achieved a test AUC of .78 (Figure 2). However, only LLM+BART had interpretable features (Figure 3), showing how the proposed framework can achieve top performance while significantly improving feature interpretability. We note that the alternative approach (b) which restricted model training to only the top features returned by the LLM achieved a test AUC of .75, which shows that one cannot assume the LLM-extracted prior is error free. The approach of training a GBT on a limited feature set from clinicians performed the worst, likely because it missed some predictive features.

LLM+BART

Feature	Variable Importance	Prior
B-type natriuretic peptide	0.64	3.0
Complications of acute myocardial infarction	0.37	3.0
Acute and unspecified renal failure	0.35	3.0
Dph erx general acetaminophen-watch meds	0.27	1.0
Erx general all flushes except heparin	0.22	1.0
Dph chf readmission laxatives	0.20	2.0
Dph chf readmission anticoagulants	0.19	3.0

GBT

Feature	Variable Importance	Prior
Erx general all flushes except heparin	0.13	1.0
B-type natriuretic peptide	0.05	3.0
Dph erx general acetaminophen-watch meds	0.03	1.0
Int'l normal'd ratio	0.02	1.0
Bun	0.02	1.0
Dph chf readmission laxatives	0.02	2.0
Erx general hydrations for infusion charge administration calculation	0.02	1.0

Figure 3: Top features for LLM+BART and GBT. Clinicans felt that the LLM+BART features were more interpretable, as citing that the selected features were more directly linked to readmission risk. Note: BART and GBT measure feature importances on different scales

Discussion This work is the first to demonstrate that Bayesian ML models with clinical priors obtained from an LLM can significantly improve the interpretability of tabular models while preserving model performance. In particular, this framework helped revise a model intended for real-world deployment at ZSFG by guiding it to rely on more clinically interpretable features. Interestingly, this framework was effective even though the priors were generated by a general-purpose LLM that was not trained on medical text. Future directions include exploring the use of clinical LLMs and other ways to encode LLM knowledge into BART priors.

References

1. Choi K, Cundy C, Srivastava S, Ermon S. Lmpriors: Pre-trained language models as task-specific priors. arXiv preprint arXiv:2210.12530. 2022 Oct 22.
2. Hill J, Linero A, Murray J. Bayesian additive regression trees: A review and look forward. Annual Review of Statistics and Its Application. 2020 Mar 7;7:251-78.
3. Flint KM, Allen LA, Pham M, Heidenreich PA. B-type natriuretic peptide predicts 30-day readmission for heart failure but not readmission for other causes. Journal of the American Heart Association. 2014 Jun 10;3(3):e000806.