

### Abstract

Motivation : Electronic health records (EHR) are widely used by many hospitals to store and organize patient information; however, the crucial information is usually buried within the extensive descriptive text. To fully exploit the utility of EHR, natural language processing (NLP) may aid doctors to summarize the patient history and status.

**Methods** : Given EHR annotated with coronary artery disease (CAD) risk factors, data was cleaned to unify the structure of every EHR. Three models: rule-based, deep learning and traditional machine learning method were compared for their performance then Naive Bayes algorithm and rulebased algorithm are combined and implemented to group each word in text into categories. Specifically, rule-based algorithm focused on family history while Naive Bayes is applied to the rest of the categories.



Fig.1 General method for modeling and user interface

**Results** : The evaluation is done on a document level to reflect whether a patient shows any signs of this risk factor. The weighted F1 score of the combined model is 0.916. The result is summarized in a user interface.

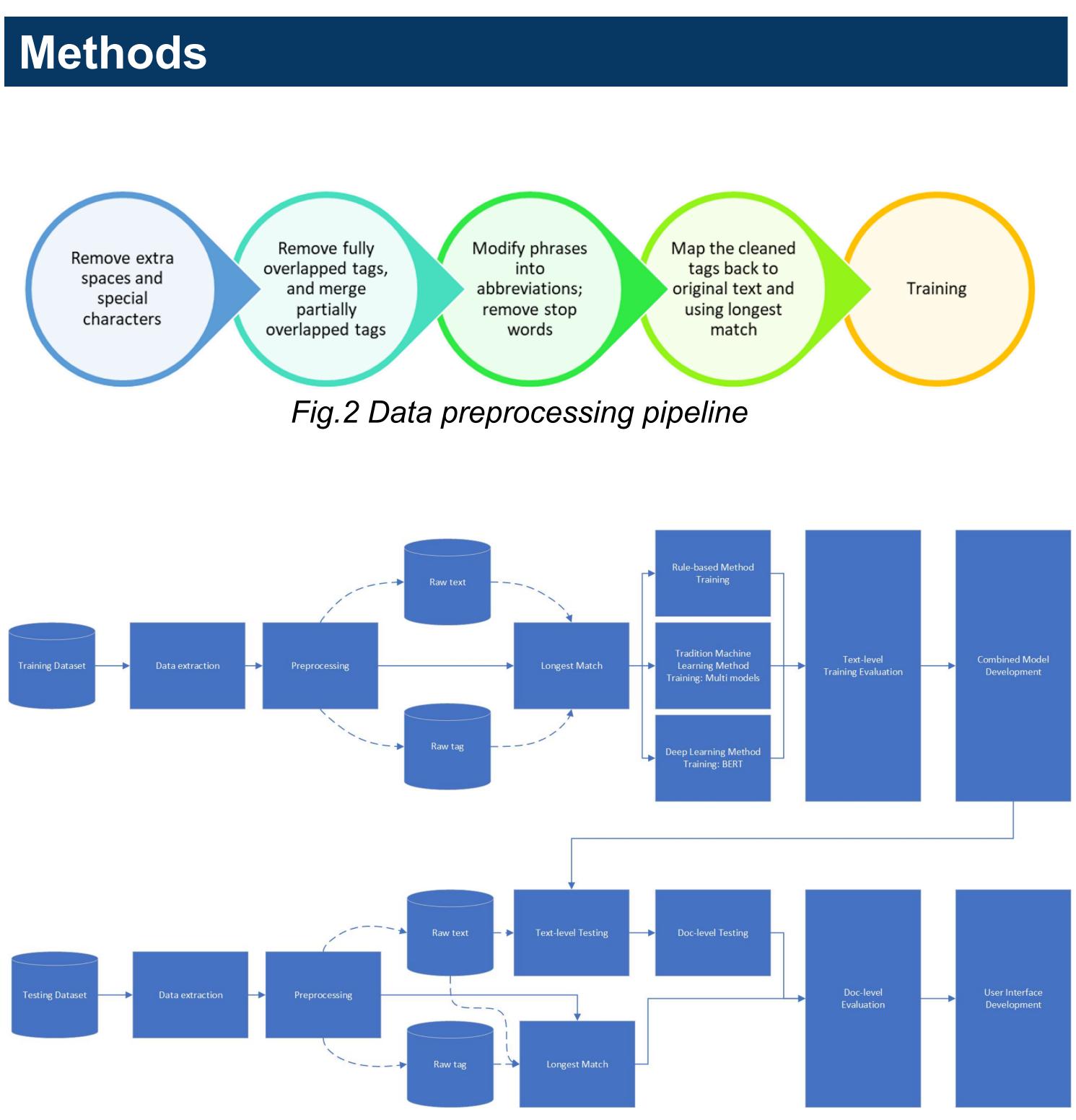


Fig.3 Detailed pipeline for modeling and user interface

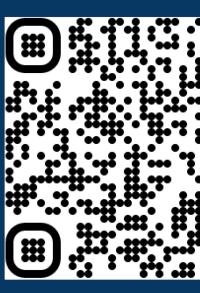
## Coronary Artery Disease Risk Factor Summarization with Longitudinal Electronic Health Record Data Todd Hartman, Yunzhou Liu, Ziyang Xu, Michelle Nguyen, Adam Charles Johns Hopkins University, Department of Biomedical Engineering

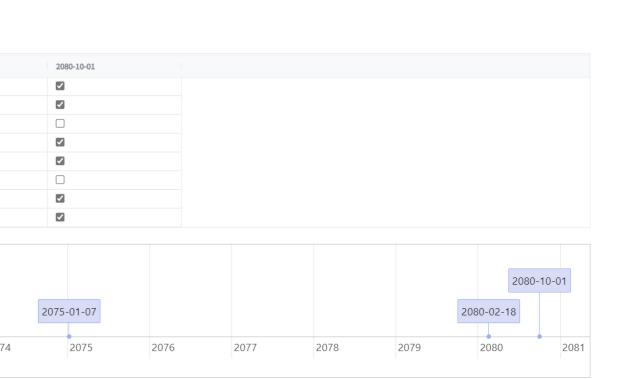


# The coronary artery disease (CAD) risk factors recognition model and summarization tool aims to track patients' risk factors over time and proposes a system for clinicians to manage patient care with reduced burden.

• (	Use the menu at	left to select a pa	tient			
folder • F	Patient's summa	ary will appear bel	ow			
o file here						
r file • CSV Cate	egory	2069-04-07		2075-01-07	2080-02-18	
CAE	)					
	DICATION					
SMO	OKER					
HYF	PERTENSION					
DIA	BETES					
FAN	MILY_HIST					
▼ OBI	ESE				$\checkmark$	
HYP	PERLIPIDEMIA					
	2069	2070	2071	2072	2073	20
He		ord: 2080-1		nal pain. History	of Present Illness	N er his b

Fig.4 User Interface of Risk Factor Summarization System Ver 1.1





CAD x 4 in . He presents with pain over the lower part of his chest since east bone, and is present all the time, worse with coughing, sneezing, or sudden deep breath. No change with bblem. No variation with food. No change with position (e.g., leaning forward). Denies productive cough, fever, t remind him of his sx prior to CABG CAD (which he says was exertional SOB), but does feel like his incisional e negative in detail. He says his pain is currently gone, and he feels fine lying back in the exam room. Problems ARPAL TUNNEL SYNDROME ?RT CAROTID DISEAS 2 Medications AMIODARONE 200 MG (200MG TABLET take 1) PO QD Aspirin MEDICATION (ACETYLSALICYLIC Maleate 10 MG (10MG TABLET take 1) PO QD x 30 davs Hctz (ATORVASTATIN) 80 MG (80MG TABLET take 1) PO OHS METFORMIN din Es 7.5/750 (HYDROCODONE 7.5mg + Apap 750mg) 1 TAB PO BID PRN . Mailing hard copy to you, Xanax 6 TEMPERATURE 98.7 F O2 SAT 97 BLOOD PRESSURE 135/60 Right arm. BLOOD PRESSURE 132/60 L arm. Physica cous membranes. Neck No visible JVD. No cervical nodes. Chest Lungs clear, good air prox 2-3 cm above the xiphoid-- subcutaneous tissue feels a little thicker here, but no fluctuance, erythema systolic murmur, no rub. ?S4. Abdomen Soft, ND, NT. + bowel sounds. Extremity Warm. No edema. Lab Tes verted Ts in II. III. AVF. Compared with 8/79-- the QRS complex is slightly narrower. Impression Sternal pain-- non-



## **Evaluation**

Performance of three different methods on text-level were tested. Deep learning method is less suitable due to the limited data source and was not used. Among classic machine learning models, Naïve Bayes shows better performance and rule-based method is applied to category with few instances.

Machine learning Deep learning Rule-based											
type	precision			recall			f1				
Micro ave	0.8990	0.6689	0.1941	0.8990	0.6470	0.777	0.8990	0.6578	0.3106		
Macro ave	0.8832	0.3872	0.2272	0.7909	0.2935	0.7706	0.8197	0.3239	0.3509		
Weighted ave	0.8996	0.6487	NA	0.8990	0.6470	NA	0.8973	0.6439	NA		

The final model evaluation is done on the document-level to focus on patient-level statistics. Medication has the best performance due to its unique word bank. CAD is less ideal perhaps due to original tag quality. However, the performance of each category has well supported the document-level risk factor prediction based on patient's health record.

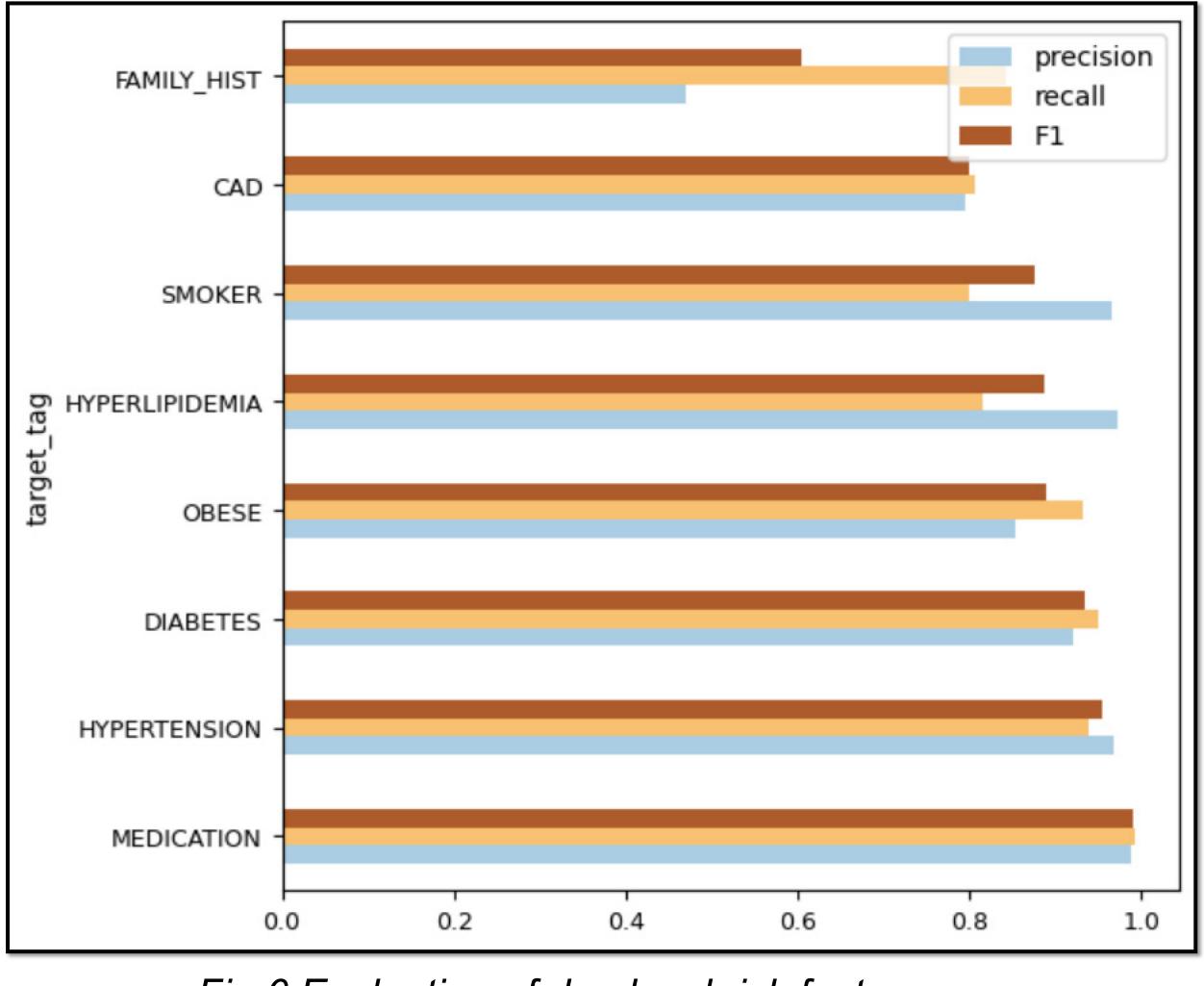


Fig.6 Evaluation of doc-level risk factors

### Conclusion

The performance of risk factor prediction is improved by using a multi-model approach. The development of document-level model and user interface may help summarize patients' risk factors over time and help clinicians manage patient care with reduced burden. In future work, we would aim to improve performance of identifying these risk factors by incorporating additional rules. We would also want to get clinician input on the utility of this tool and design feedback.

Fig.5 Evaluation and Comparison of different models