Breakout Session 1: Track A

Empowering Cloud Computing for Non-image-based Diabetic Retinopathy Screening by Designing an EHR-oriented Incremental Learning Framework

Dr. Tieming Liu

Professor, Oklahoma State University

NOT-OD-23-070: Empowering Cloud Computing for Non-image-based Diabetic Retinopathy Screening by Designing an EHR-oriented Incremental Learning Framework

Chenang Liu (co-I), Tieming Liu (PI)
School of Industrial Engineering and Management
Oklahoma State University

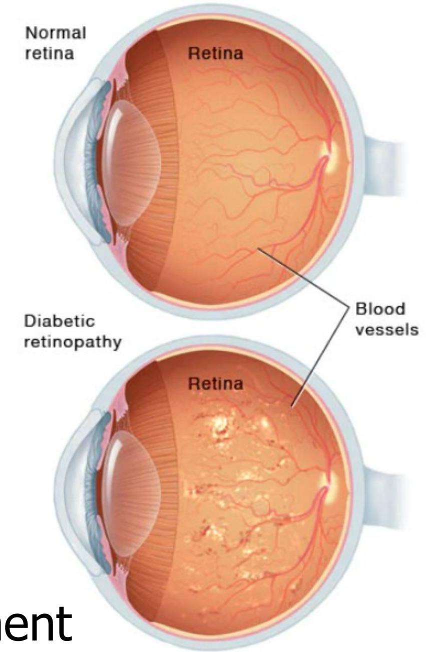
chenang.liu@okstate.edu, tieming.liu@okstate.edu

Parent Grant: NIH-NEI: 5R01EY033861
Harnessing Tensor Information to Improve EHR Data Quality for Accurate Data-driven Screening of Diabetic Retinopathy with Routine Lab Results

Motivation

Diabetic Retinopathy (**DR**)

- Most common cause of vision loss among diabetic patients
- Leading cause of blindness among adults in developed countries[1]
- 7.69 M (2010) to 14.6 M (2050) in U.S.[2]



- **Early stages**: unsymbolic and most effective period for treatment
- **Low compliance rate** (~43%) for recommended annual eye exams

1, T. A. Ciulla, A. G. Amador, and B. Zinman, "Diabetic retinopathy and diabetic macular edema: pathophysiology, screening, and novel therapies," Diabetes care, vol. 26, no. 9, pp. 2653–2664, 2003.
2, National Eye Institute, NIH. Diabetic Retinopathy Data and Statistics. https://www.nei.nih.gov/learn-about-eye-health/outreach-campaigns-and-resources/eye-health-data-and-statistics/diabetic-retinopathy-data-and-statistics. Updated on 11/19/2020

Problem Statement

Current Screening Method

- Annual eye exams
  - Lack of experts
  - Dilation
  - Cost
- AI-based retinal imaging method
  - Expensive imaging equipment

Our approach:

- non-image based Screening
  - Lab test data (widely available)
  - Using non-temporal data
  - Using temporal data

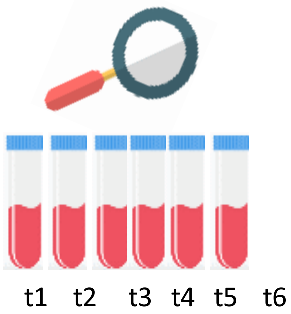
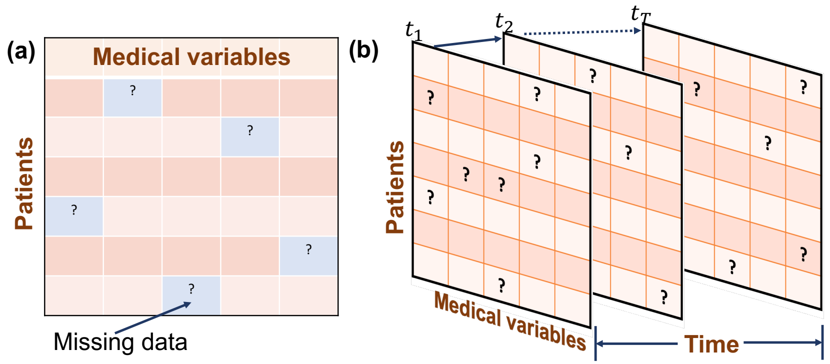


Image sources: yoursightmatters.com; Carl Zeiss

Aims of Parent Grant



Technical Challenges	Harnessing Tensor Information to **Improve EHR Data Quality**
• **Missing Data**	• Aim 1: weighted K-Nearest Neighbors (wKNN) for data imputation
• **Imbalanced Data**	• Aim 2: augmented generative adversarial network (GAN) for data balancing
• **Unlabeled Data**	• Aim 3: Bayesian hierarchical modelling for classifying unlabeled patients
• **Tensor Data**	• Aim 4: Multi-branching Temporal Neural Networks for disease prediction

Data and Variables

Cerner Health Facts®

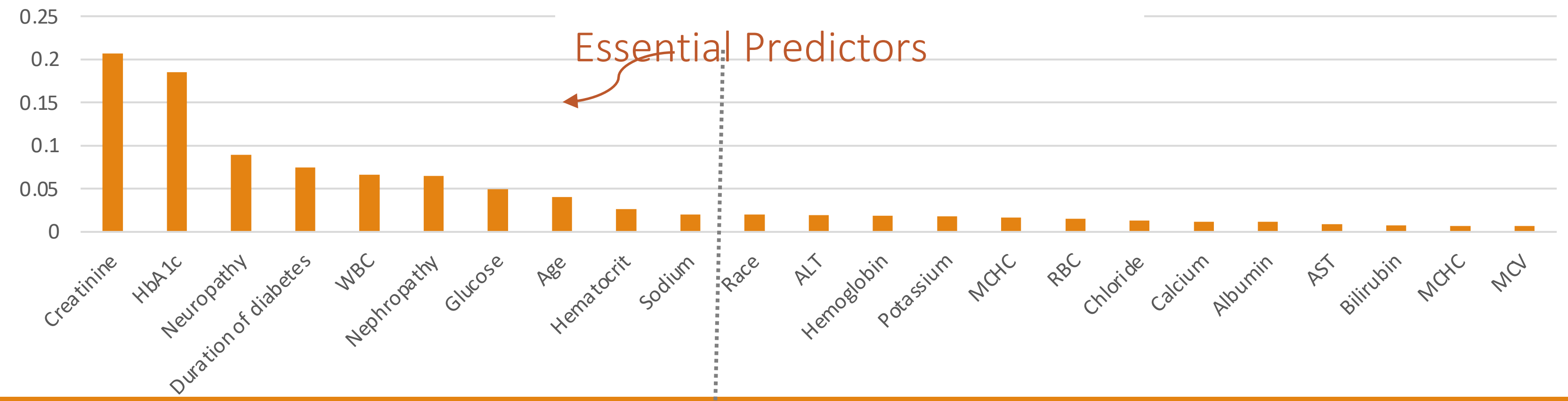
EHR Database

- Patient #: > **100.8** M
- Span: since 1998

	# of **DR** Patients	# of **Non-DR Diabetic** Patients	Positive Rate
Original Dataset	69,354	2,363,051	2.85%
Final Dataset (with >=**10** records)	12,590	401,609	3.04%

Independent Variables:

- 21 common lab tests
- 3 demographics (race/gender/age)
- 5 comorbidities



Opportunities and Challenges

Opportunity:
- Cerner moved to the Cloud
- Periodically updated database

Challenges:
- Simply retraining the model with all the date will result in an extremely high computational burden on the cloud.
- Need an efficient and effective model update approach

Approach: *Incremental Learning* (IL)

Formulated incremental learning problem for this project
- Update the model by integrating the new data and the existing model, mathematically,

$$f' = \mathcal{G}(f, \boldsymbol{\mathcal{Y}} \backslash \boldsymbol{\mathcal{Y}}')$$

- $f(\cdot)$ is DR prediction model, and $f'(\cdot)$ is the updated prediction model by incorporating new EHR data $\boldsymbol{\mathcal{Y}} \backslash \boldsymbol{\mathcal{Y}}'$ using IL framework $\mathcal{G}$. $\boldsymbol{\mathcal{Y}}'$ is the updated data, and "\" represents set subtraction.

Aim 1: Design an EHR-oriented IL Framework

Motivation & Gap

- An EHR-oriented IL framework for DR prediction is still unavailable.
- Most of the state-of-art IL approaches do NOT meet the need of:
  - Preserving previously acquired knowledge
  - Considering the longitudinal effects in EHR

Proposed Approach

*A sample recycling-assisted incremental learning (**SR-IL**), which*

- *partially access the existing dataset via adaptive sampling strategy*
- *reduce the potential information loss*

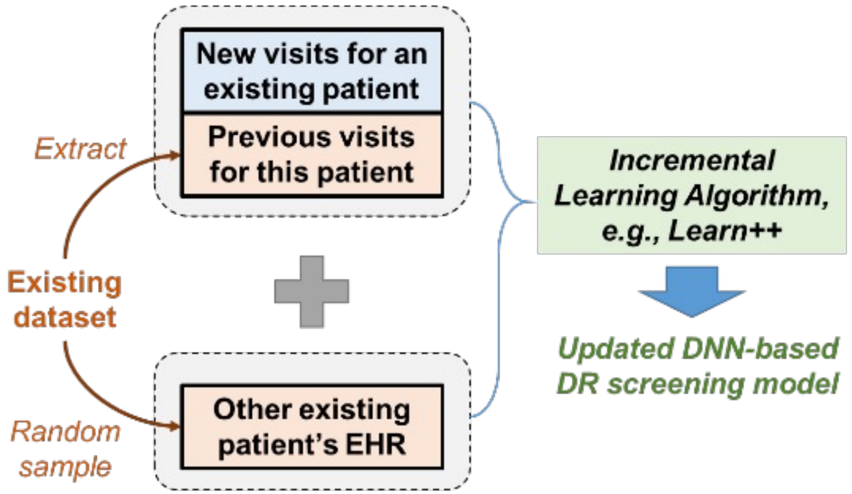


Figure: The overall framework of the proposed SR-IL.

Current Progress: A Preliminary Study

Promising results:

- Assisted by importance (***give higher weight to the DR samples***) sampling, the proposed approach has the lowest false negative and true positive occurrence.

Classifier	False negative IL	False negative IL SS	False negative IL IS	False negative CL	True positive IL	True positive IL SS	True positive IL IS	True positive CL
Logistic Regression	1136	1082	342	1013	114	168	908	237
Decision Tree Classifier	1013	889	445	798	237	361	805	452
Random Forest Classifier	1240	1199	263	1035	10	51	987	215
Gradient Boosting Classifier	1150	1076	274	959	100	174	976	291
AdaBoost Classifier	1070	954	329	932	180	296	921	318
Extra Trees Classifier	1244	1216	302	1072	6	34	948	178
Hist Gradient Boosting Classifier	1151	1090	288	945	99	160	962	305
SVC	1250	1238	301	1024	0	12	949	226
Gaussian NB	875	754	546	781	375	496	704	469
MLP Classifier	1060	956	298	835	190	294	952	415
Gaussian Process Classifier	1126	1088	344	992	124	162	906	258
Quadratic Discriminant Analysis	1053	615	1014	239	197	635	236	1011
Linear Discriminant Analysis	1039	978	343	935	211	272	907	315

- "IL" – Incremental Learning without sampling,
- "IL SS" – Incremental Learning with Simple Sampling,
- "IL IS" – Incremental Learning with Importance Sampling,
- "CL" – Traditional (Classic) Machine Learning.

Aim 2: Scale-up IL to the Cloud Platform

Goals & Plan

- Make the implemented SR-IL toolbox ***compatible*** with the cloud computing platform, which requires
  - Effective integration of programming codes
  - Appropriate adoption of the dependent computing toolboxes and their versions

- Scale up and test the performance of ***SR-IL for large-scale EHR dataset***, including both
  - Computational efficiency
  - DR risk prediction accuracy

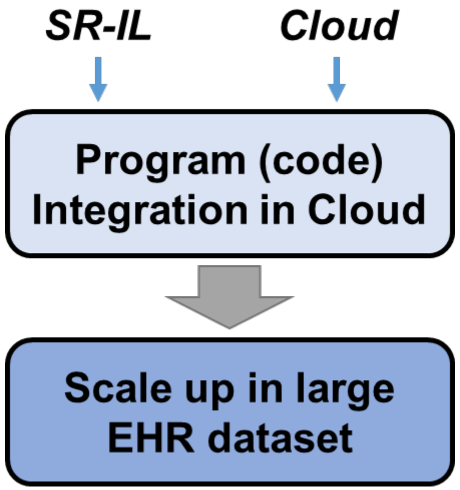


Figure: The overall procedures of Aim 2.

Testbed Platform

There will be two testbed platforms
- A local testbed
- The AWS cloud testbed

	Validation for computational efficiency	Validation for prediction accuracy
Evaluation Metric	Actual computational time	AUC score or recall score
Benchmark	(1) Direct DNN model retrain without IL; and (2) Common IL approaches;	
Data Used	Cerner Real-World Data (CRWD)	
Criterion for Success	Compared to the benchmark (2), SR-IL's computational efficiency is comparable, and the prediction accuracy is much better.	Compared to the benchmark (1), SR-IL's prediction accuracy is comparable , and computational efficiency is much better.

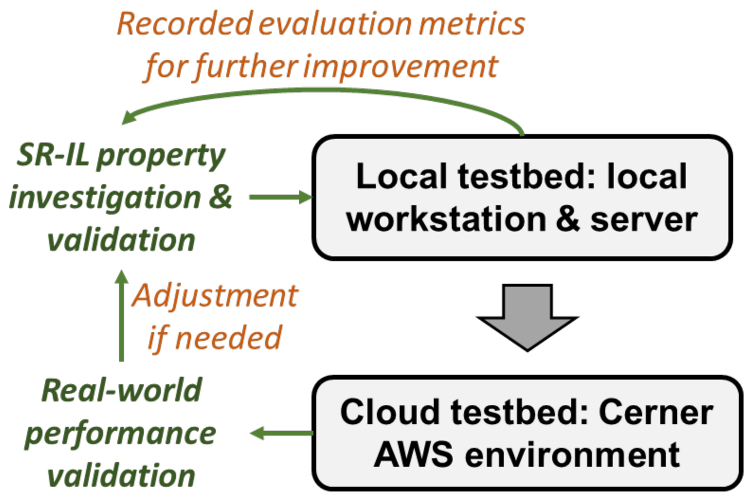


Figure: Illustration of our testbed and evaluation & validation plan.