신윤열 (Yunyol Shin)

Researcher, Developer

Portfolio

Profile

Cypress Bay Highschool, FL, USA

KAIST, Undergrad, Computer Science

Seoul National University, Master's Degree, Computer Science and Engineering

Master's Degree

- MS degree from Sun Kim's Lab in Computer Science and Engineering Departments, SNU
- Research with Seoul National University Hospital

Interests/Knowledge

- Out-of-Distribution Detection, Natural Language Processing, Deep Learning

Main Research Theme

- Application of Uncertainty Estimation and Calibration in Medical Domain

Skills

- Python, C++, Java, R, Spring Framework, Android, Unity, MySQL, MongoDB, JavaScript, HTML, CSS, PyTorch

Language

- Korean, English(TOEIC 990)

M.S. in Computer Science, SNU
Bio&Health informatics Lab.
E-mail: yunyol@snu.ac.kr
Mobile: 010-8725-2571
Address: 4-Dong 103-Ho, Olympic-Ro 4-gil 15, Songpa, Seoul

Table of Contents

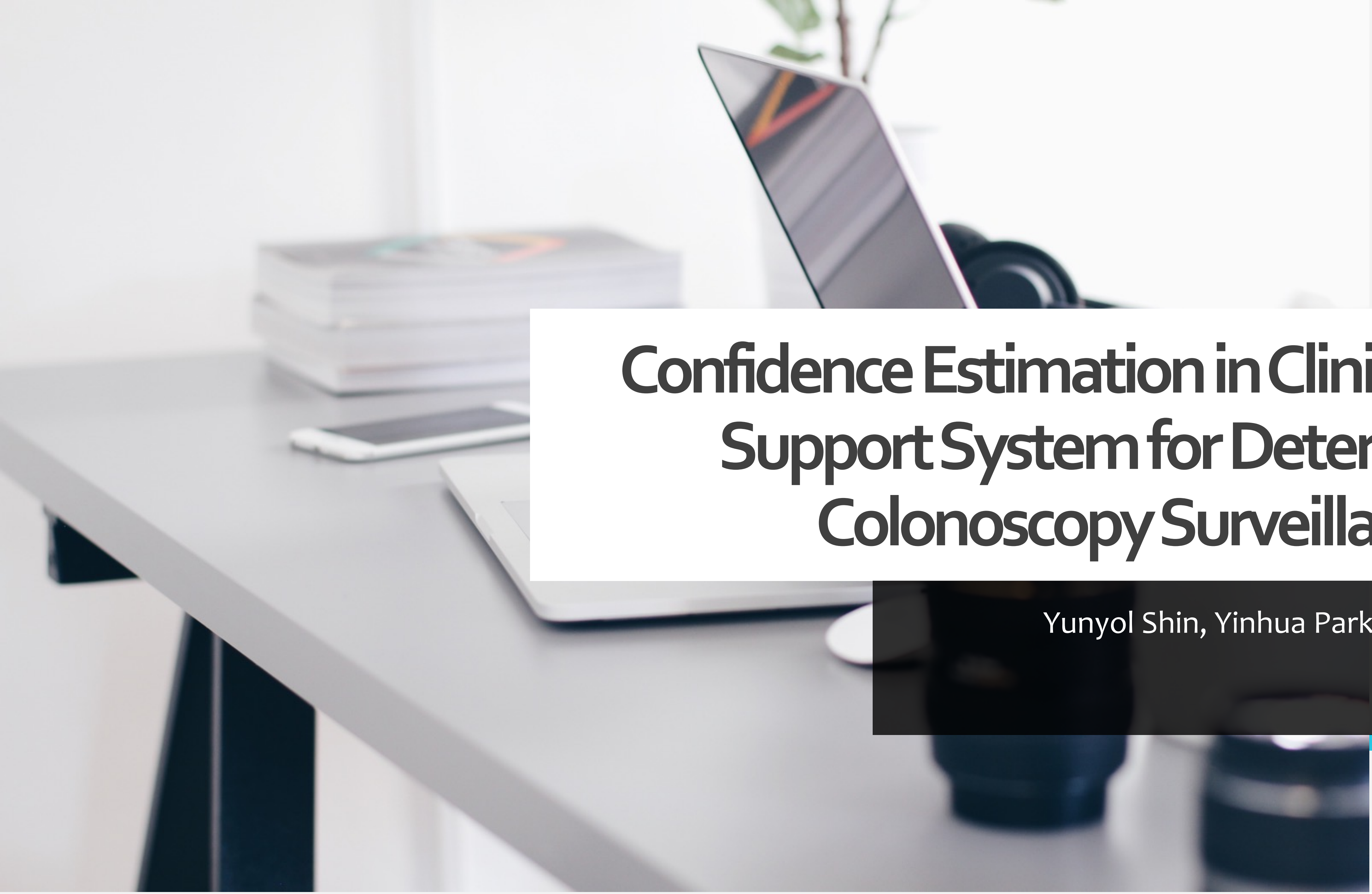
Research Experience, Develop Experience

Research Experience

- Confidence Estimation in Clinical Decision Support System for Determination of Colonoscopy Surveillance Interval (First Author)
- Daily Appointment Non-Show Prediction with Calibrated Neural Network (First Author)
- COVID-19 Virus Whole Genome Embedding Strategy through Density-based Clustering and Deep Learning Model (Co-Author)
- AutoCoV: Tracking the Early spread of COVID-19 in Terms of the Spatial and Temporal Dynamics from Embedding Space by K-mer Based Deep Learning (Co-Author)

Develop Experience

- Projects
  - Unity
    - Multi-user VR racing game on Android using Oculus
    - A simple reach-the-destination ball game by flipping environment 90 degrees
  - Android
    - Webtoon platform crawling and showing webtoons from Daum, Naver, Lezhin
    - Multi-user radio app where users can add songs to the playlist
- Developer at Balance Hero
- Teaching Experience at Sparta Coding Club and AAiT(Volunteer)



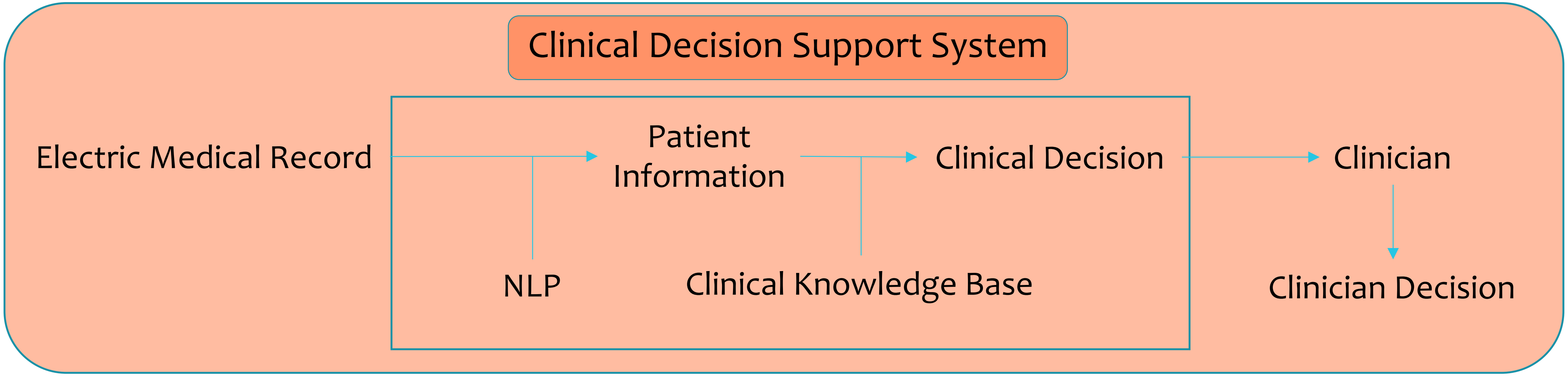
Confidence Estimation in Clinical Decision Support System for Determination of Colonoscopy Surveillance Interval

Yunyol Shin, Yinhua Park, Sun Kim, Jungho Bae

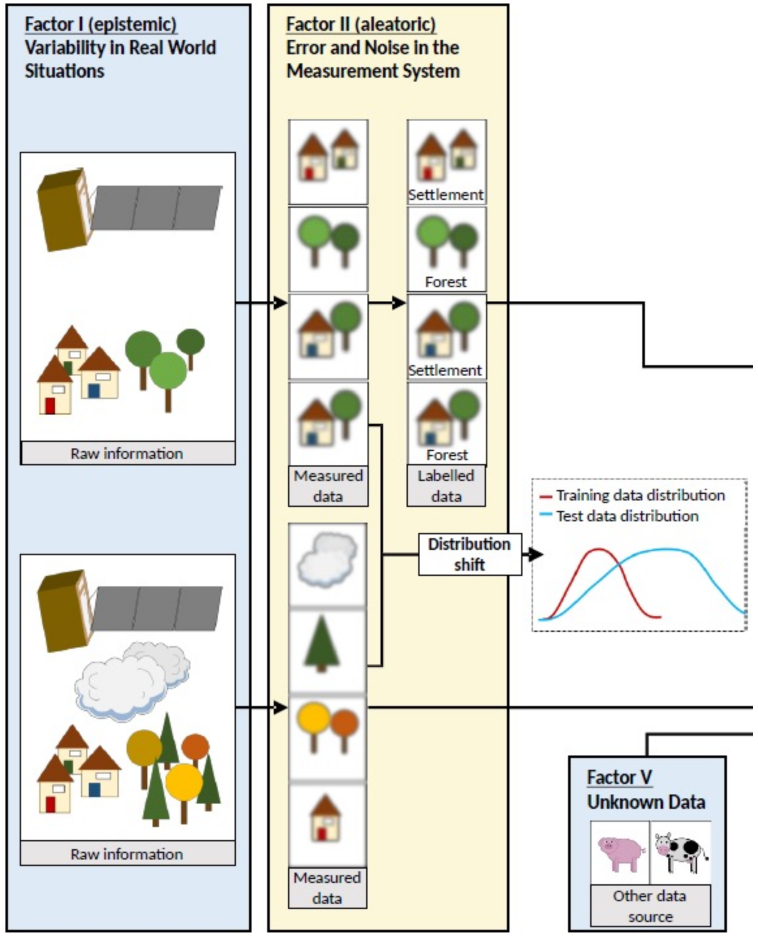
Background

- What is Clinical Decision Support System (CDSS)?
  - Intended to improve Healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information

- CDSS limitation
  - Is not Robust on errors outside Training Data



Background

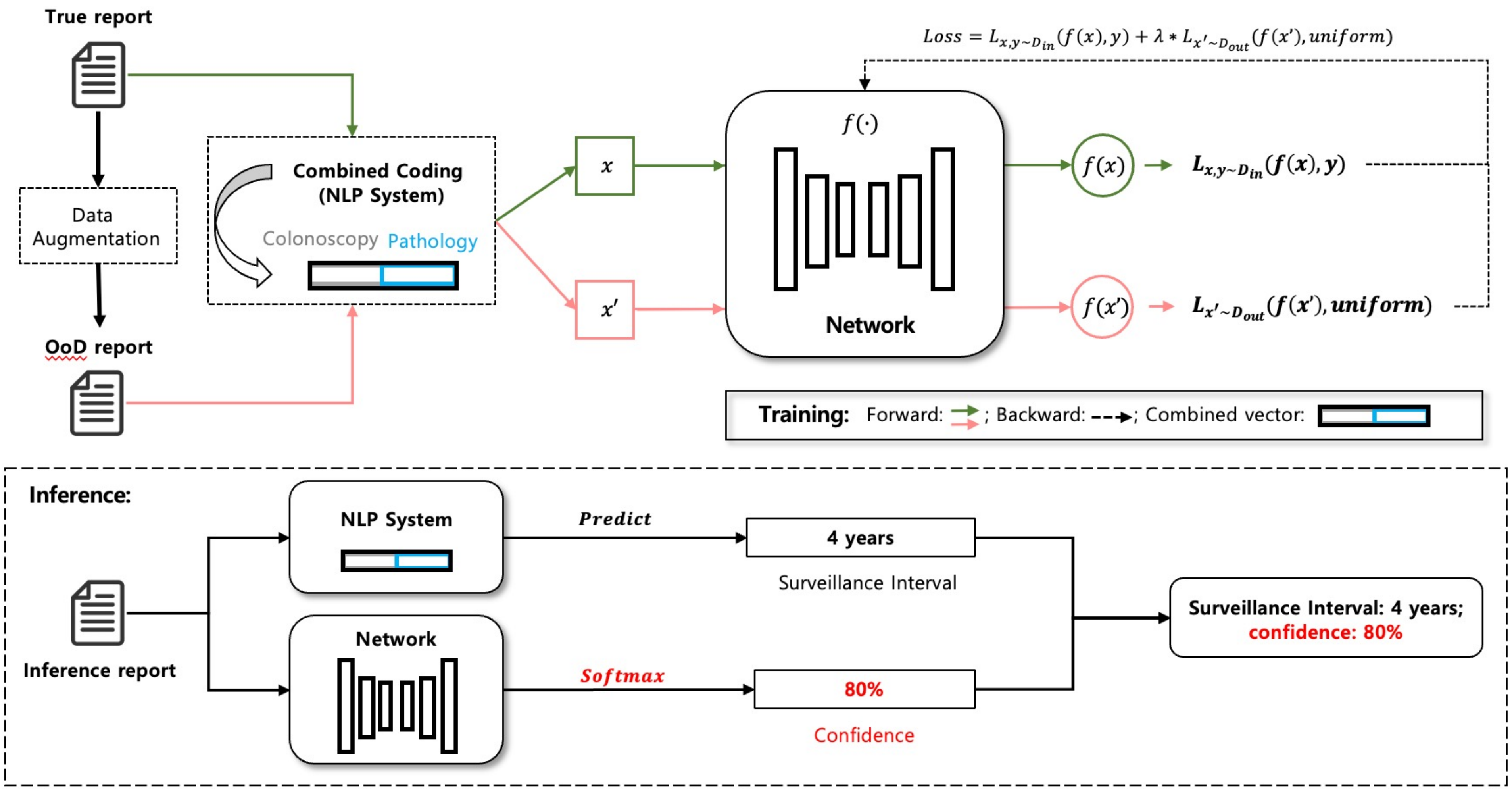


Jakob Gawlikowski et al. arXiv 2021

- Data Uncertainty
  - Uncertainty in the system's prediction caused by incomplete collection of data

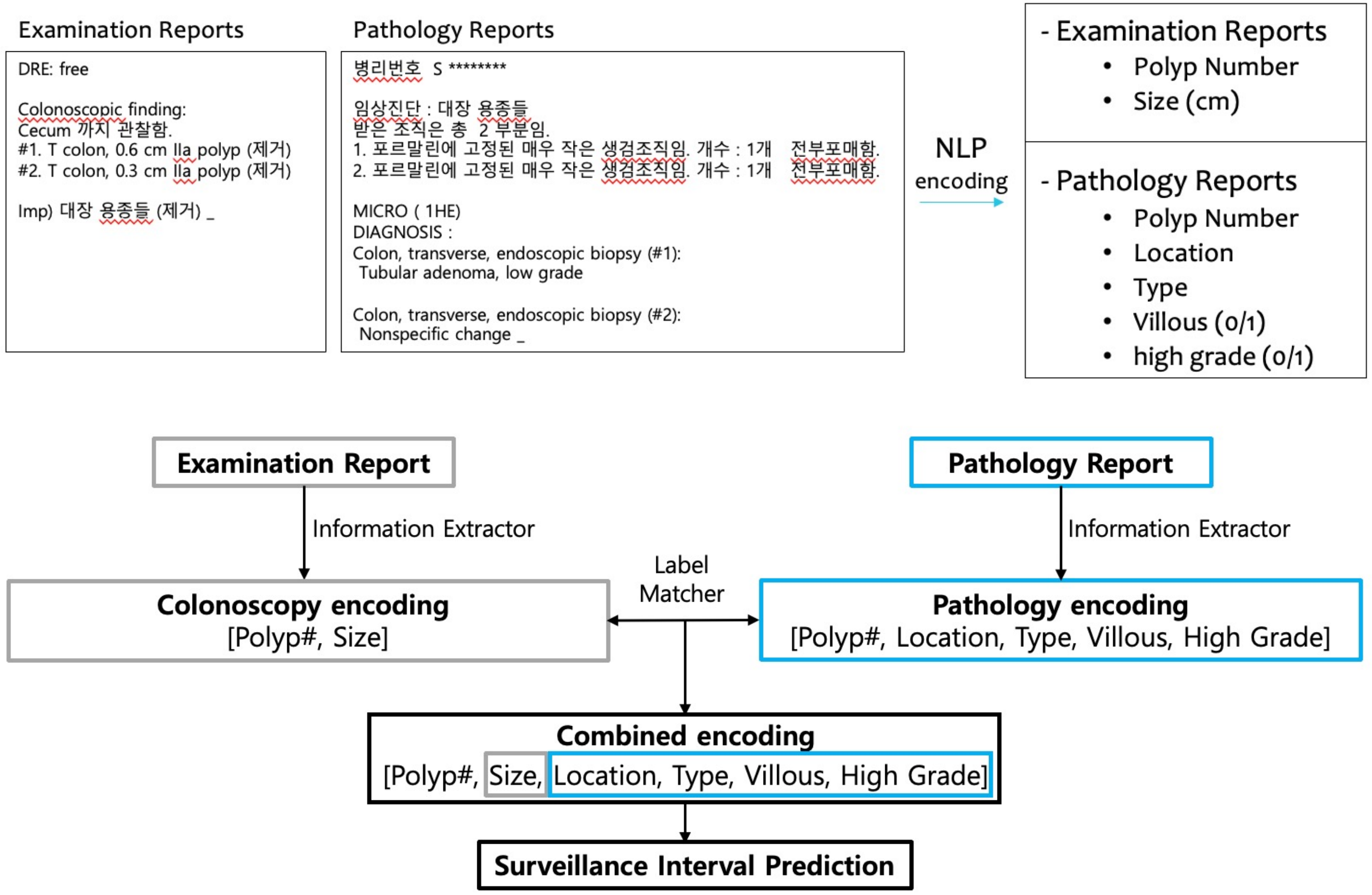
- Uncertainties in Electronic Medical Record(EMR)
  - Epistemic (Systematic, Domain) Uncertainty
    - Each institution cannot collect data characteristic to clinicians in the whole world
    - Cannot reflect all staff changes over time
  - Aleatoric (Statistical, Random) Uncertainty
    - Cannot collect all possible erroneous reports due to typos, or structural violations

Overall Model Structure



Misclassified Dataset is generated from Original Training Data. Original data and augmented data are used for NN model training.

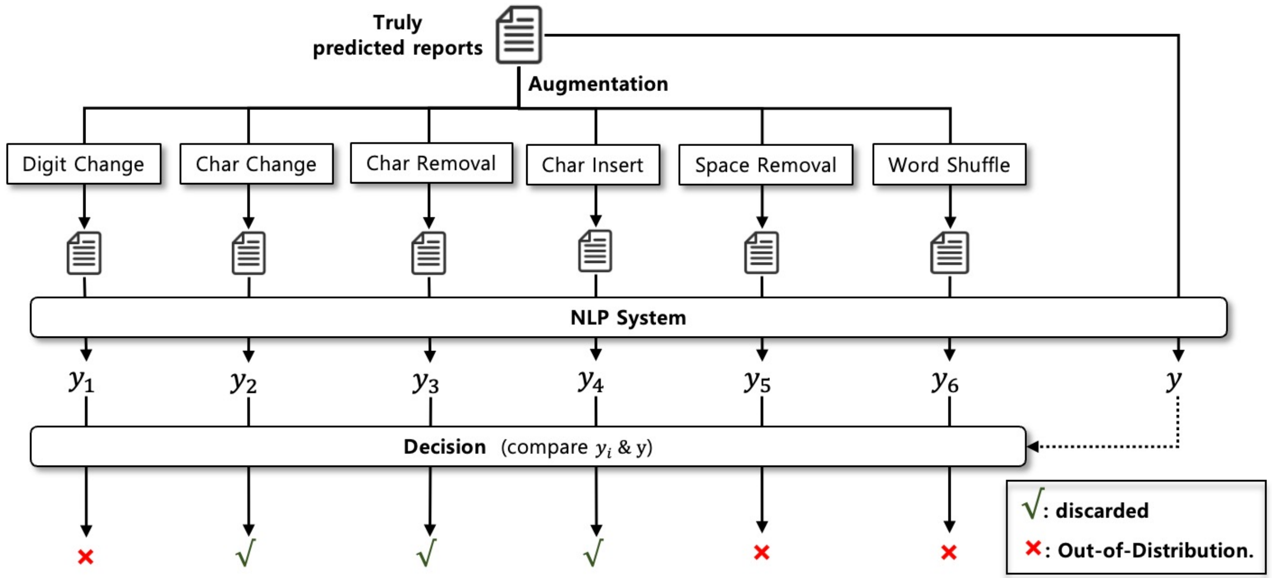
NLP System



Locations	Type
• 0: IC Valve • 1: Cecum • 2: Ascending • 3: Transverse • 4: Descending • 5: Sigmoid • 6: Rectum • 7: Terminal ileum • 8: Appendix	• 0: No type specified • 1: Hyperplastic • 2: Adenoma • 3: Serrated adenoma • Subtype not specified • 4: Traditional serrated adenoma • 5: Sessile serrated adenoma • 6: Carcinoid/Neuroendocrine tumor • 7: Carcinoma

The NLP system. Information extractor extracts polyp information separately from Examination Report and Pathology Report. Then polyp encodings are compared to see if extracted information matches. Surveillance Interval is predicted from combined encoding.

Neural Network Training



	Examination Reports		Pathology Reports	
	Generated	Used	Generated	Used
Digit Change	37	37	206	206
Character Change	549	529	16681	600
Character Removal	2886	600	15016	600
Character Insert	16598	600	2886	600
Space Removal	293	293	10	10
Word Shuffle	567	567	10633	600

- Used character-wise Models
  - word-wise models cannot take typos or structural violations into consideration
  - 각각의 글자를 one-hot encoding 하여 사용
- Model
  - 1-d Convolutional Network (ResNet)
  - Outlier Exposure

$$\text{Loss} = L_{x,y\sim D_{in}}(f(x), y) + \lambda * L_{x'\sim D_{out}}(f(x'), uniform)$$

Result 1- NLP System Accuracy

Method	Accuracy
Clinician Label + 1 Self Validation	99.3%
NLP System - Domain-Specific	99.9%
NLP System - cTAKES-based	94.1%

	2013	2014	2015	2016	2017	2018	2019	2020
# Exam/Path Reports	4013	4126	5678	6428	6186	6132	6148	5692
# Mismatched Reports	259	220	191	110	92	91	100	65
Mismatch Ratio	6.45%	5.33%	3.36%	1.71%	1.49%	1.48%	1.63%	1.14%

Result 2- Error case Detection

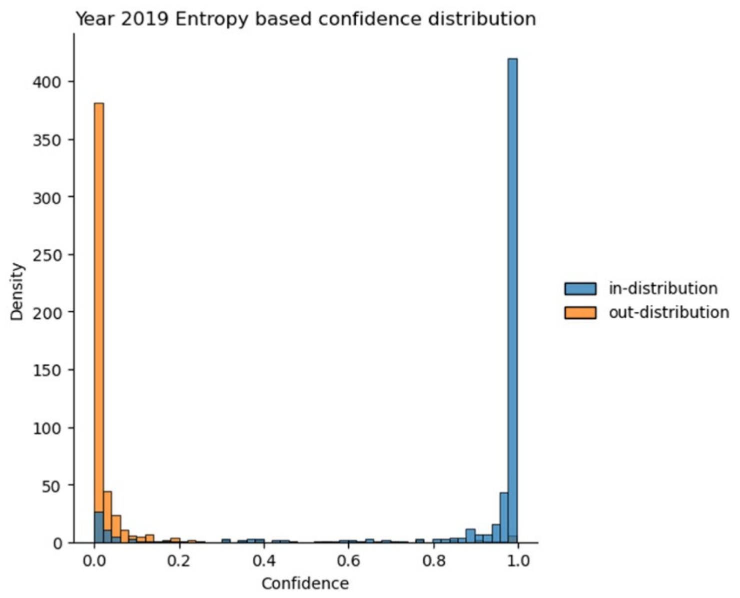
	AuPR (MSP)	AuPR (NSE)	Accuracy (@threshold)	Average NSE
MLP	0.6859	0.6871	0.5755 (@t=0.89)	0.7108
- After OE	0.7533	0.7551	0.6461 (@t=0.31)	0.1138
RNN-based	0.5426	0.5434	0.5487 (@t=0.40)	0.4752
- After OE	0.6290	0.6416	0.6086 (@t=0.09)	0.0802
BERT	0.6928	0.6864	0.5934 (@t=0.71)	0.6872
- After OE	0.7466	0.7474	0.6318 (@t=0.26)	0.1259
CNN	0.7686	0.7708	0.7337 (@t=0.97)	0.6946
- After OE	0.9748	0.9748	0.9303 (@t=0.23)	0.0658

$$NSE = 1 - \frac{H(\hat{y})}{H(uniform)}$$

More Findings

1. Manual Review of Reports from 2013~2015 with top 400 confidence values and bottom 400 confidence values

- Among those with bottom 400 confidence estimates
  - 9 incorrectly identified samples
- Among those with top 400 confidence estimates
  - 0 incorrectly identified samples



2. Institutions can effectively seek out large portion of erroneous data by reviewing small number of reports

- Below bottom 1% confidence of correctly identified data
  - 45.1% erroneous data
- Below bottom 10% confidence of correctly identified data
  - 95.5% erroneous data

3. Low False In-Distribution data and higher False Out-Distribution data

- False Out-Distribution data is not as critical as false in-dsitribution data

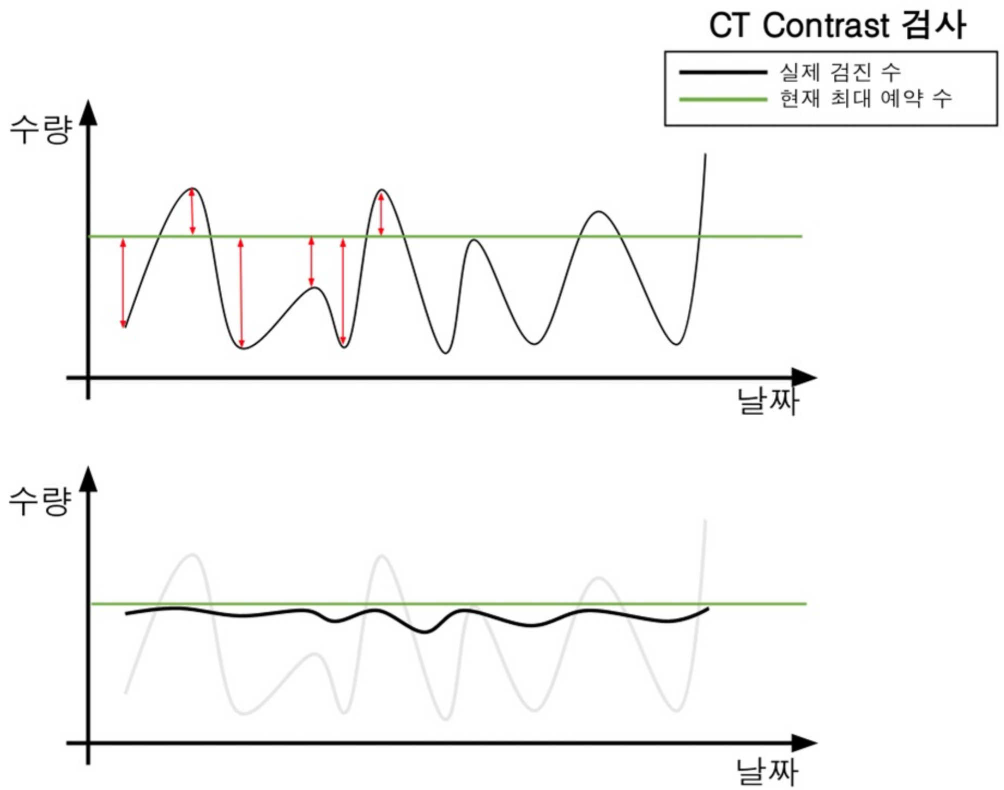
Daily Appointment Non-Show Prediction with Calibrated Neural Network

Yunyol Shin, Seungho Choi, Yinhua Park, Sun Kim, Jungho Bae, Youngah Kim, Jungmin Kim, Kyungjin Park, Youngsun Kim

Introduction

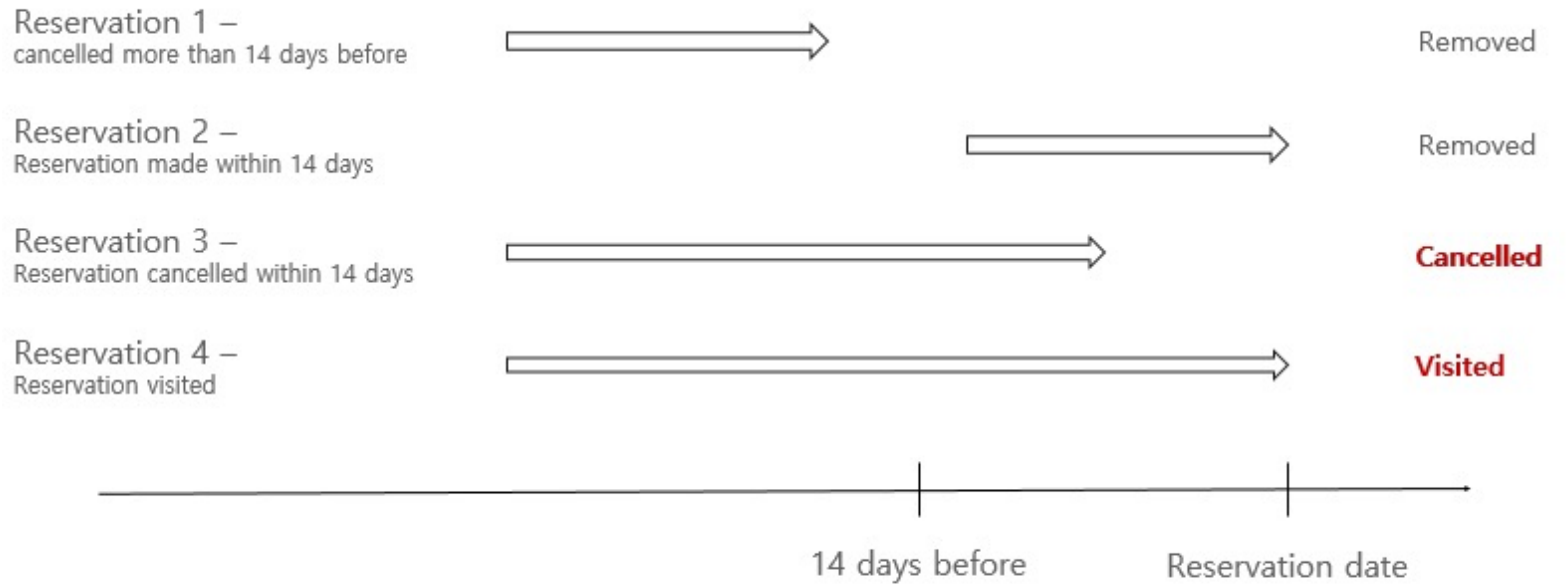
- Many Institutions receives Overbooking
  - Loss occurs when there are unused resources

- If you can predict how much of the reservations will be cancelled, you can expect to either
  - Allocate resource effectively
  - More active overbooking



Introduction

Goal: Make non-show prediction on appointment at 14 days before the appointment date!



Overbooking around 5%
Cancellation around 18 %

Introduction

Test	1-9월(평일)	2주전	10-12월(집중기)	2주전	1-9월(토)	2주전	10-12월(토)	2주전
Endoscopy	108	113	127	135	98	103	108	115
Colonoscopy	46	48	48	50	24	26	44	46
Breast SONO	50	52	60	62	35	38	45	48
OBGY	55	58	60	62	50	55	55	58
SONO ALL	190	200	240	252	180	190	210	220

- Currently, Capacity Based on…
  - Season
  - Weekday/Weekend
  - Type of Examination

- We can do better if we predict cancellations based on
  - Personal Information -> Age, Residence, VIP, Company, Cancellation History
  - Reservation Information -> How early reservation is made, Examination
  - Reservation Date -> Before/After Holidays, Weekday, Part of Month, Weather
    - Holiday Info: http://apis.data.go.kr/
    - Weather Info: https://data.kma.go.kr/

Background

- Non-show Prediction on a single Reservation date is directly related to overbooking capacity

- These models either do not make prediction on a single reservation date or predict number of non-show as sum of each patient's predicted non-show

$$\#Cancellation = \sum^{peday} I(p)$$

where

$$I(p) = \begin{cases} 1 & if\ f(x) >= threshold \\ 0 & if\ f(x) < threshold \end{cases}$$

Background

- Non-show Prediction on a single Reservation date is directly related to overbooking capacity

- These models either do not make prediction on a single reservation date or predict number of non-show as sum of each patient's predicted non-show

$$\underline{\#Cancellation} = \sum^{peday} \underline{I(p)} \quad \text{where} \quad I(p) = \begin{cases} 1 & \text{if } f(x) >= threshold \\ 0 & \text{if } f(x) < threshold \end{cases}$$

B. How do we know a group of patient will make same number of non-shows next time?

A. How can we determine that patient with given trait is going to miss the appointment? We don't know what keeps the patient from visiting!

Background

- A. How can we determine that patient with given trait is going to miss the appointment? We don’t know what keeps the patient from visiting!

- Many of no-show prediction models
  1. Make correlation analysis of features to find subset of features having high correlation with non-show rate
  2. Make logistic regression with found features to predict patient’s no-show

-> Patient no-show is determined 5-10 feature!

With this much information, there should only be appointments having a lower or higher no-show rate than mean no-show rate!

Background

- B. How do we know a group of patient will make same number of non-shows next time?
- Let's say individual patient's no-show rate follows Gaussian

$$\underset{p}{\textbf{Argmax}} \; P(\#\, Cancellations | \#\, Cancellations \sim \prod^{n} Bernoulli(p),\; p \sim N(\mu, \sigma))$$

Approximate with Binomial Distribution,

$$\underset{p}{\textbf{Argmax}} \; P(\#\, Cancellations | \#\, Cancellations \sim Binomial(n, p),\; p \sim \frac{N(\mu, \sigma)}{n})$$

p will represent most probable probability of the patients no-show rate for given number of cancellations
Using n=100, Gaussian approximation for Binomial distribution, and N(18, 20) for no-show rate distribution,

When # Cancellation = 31, p = 20.772

Real Outcome~Bernoulli & Predicted~Normal

$$\underset{p}{\textbf{Argmax}} \quad P(\#\,Cancellations \mid \#\,Cancellations \sim \prod^{n} Bernoulli(p),\ p \sim N(\mu, \sigma))$$

Approximate with Binomial Distribution,

$$\underset{p}{\textbf{Argmax}} \quad P(\#\,Cancellations \mid \#\,Cancellations \sim Binomial(n, p),\ p \sim \frac{N(\mu, \sigma)}{n})$$

Binomial is discrete Function, use Gaussian Approximation to make continuous distribution

Then the Problem Becomes,

$$\underset{p}{\textbf{Argmax}} \quad \underbrace{\frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\#Cancellation - p}{\sigma_1}\right)^2}}_{\text{Binomial Sampling}} * \underbrace{\frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{p - \mu * n}{\sigma_2}\right)^2}}_{\text{Group Cancel rate Sampling}}$$

Predicted vs. Label

Since it’s Gaussian Approximation, Differentiable!

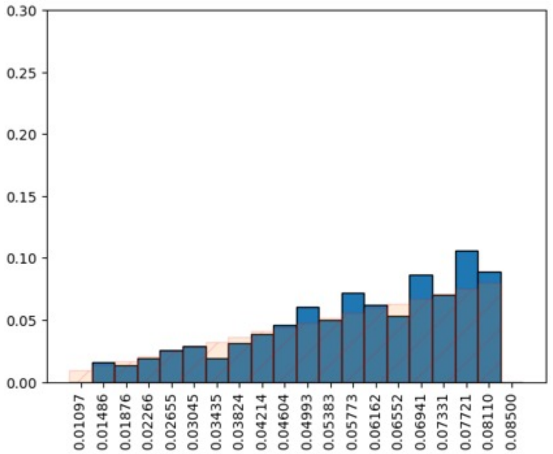
$$\underset{p}{\text{Argmax}} \underbrace{\frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\#Cancellation - p}{\sigma_1}\right)^2}}_{\text{Binomial Sampling}} * \underbrace{\frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{p - \mu * n}{\sigma_2}\right)^2}}_{\text{Group Cancel rate Sampling}}$$

To reduce computation complexity, sigma 1 is fixed.
This assumption does not change computation a lot since sigma 2 is small

$$p = \frac{\sigma_1^2 * \mu * n + \sigma_2^2 * \#Cancellations}{\sigma_1^2 + \sigma_2^2}$$

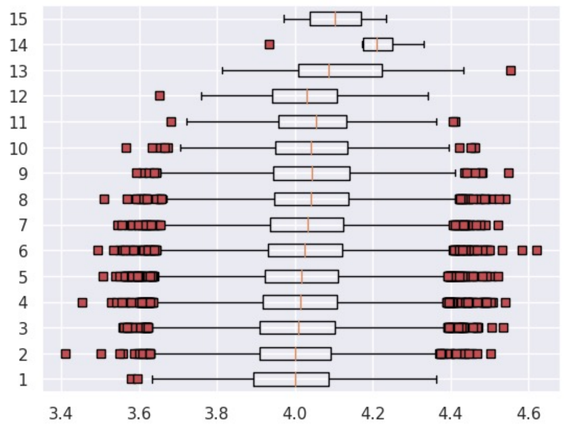
With sigma 1 fixed at mean cancellation rate,
#Cancellation = 31, then p = 20.772

Individual Patient no-show rate

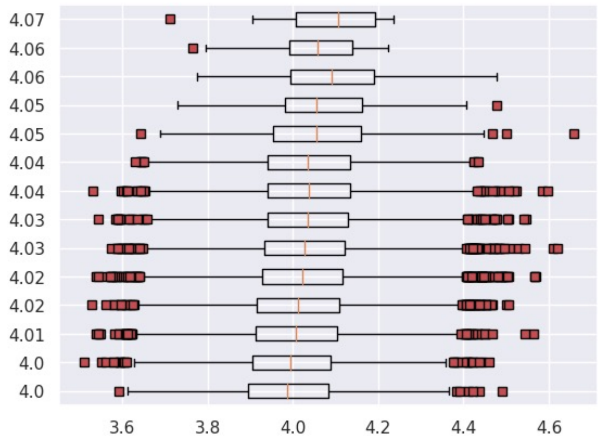


50000 random permutations of 100

\# Cancellations



Posterior Adjusted Label



Background

- Instead of making # Cancellation prediction, we find Expected # Cancellation on each day,

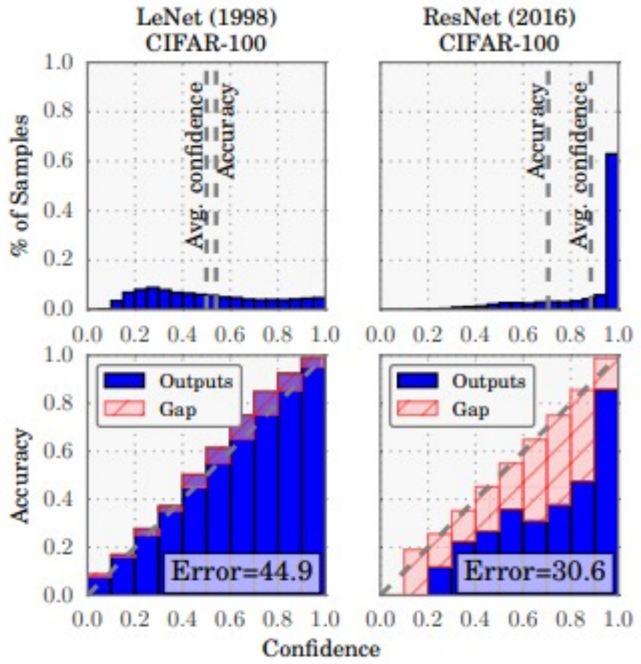
$$E[Day\ Cancellation] = E[\sum Patient\ Cancellation]$$

$$= \sum \underbrace{E[Patient\ Cancellation]}_{\text{Patient Cancellation rate}} \qquad by\ Linearity\ of\ the\ expectation$$

by summing over cancellation rate of individual patient on the day

Background

- Neural Network Uncertainty Calibration
  - Modern Neural Networks are usually Overly Confident
  - Average Confidence is much higher than accuracy



On Calibration of Modern Neural Networks; Chuan G, et al.

- Calibration Methods
  - MixUp, Label Smoothing, Batch Normalization, BNN, Regularization methods, etc..

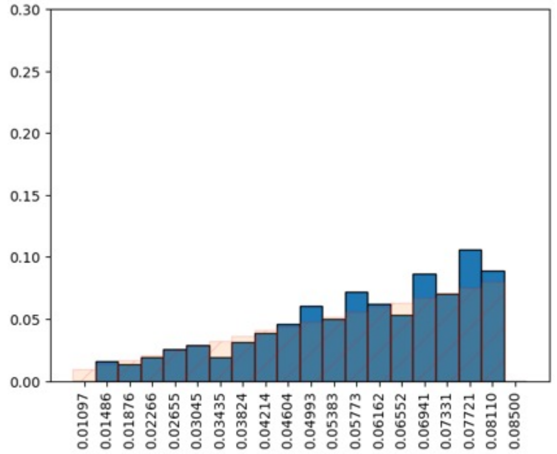
Methods

- Problem of Batch Learning

- Variation of samples selected from Bernoulli outcomes (Batch)
  - $n * p(1 - p) = 100 * 0.18 * 0.82 = 14.76$

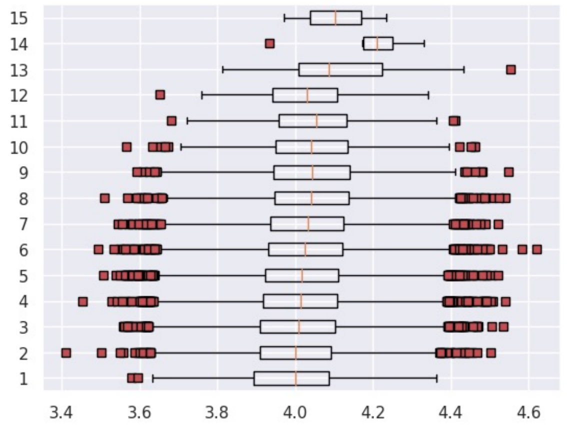
- Variation of samples selected from Gaussian Rate (Real World)
  - Even with unrealistically large std of 0.2,
  - $n * (0.2)^2 = 4$

Dummy Example with no-show rate = 4%

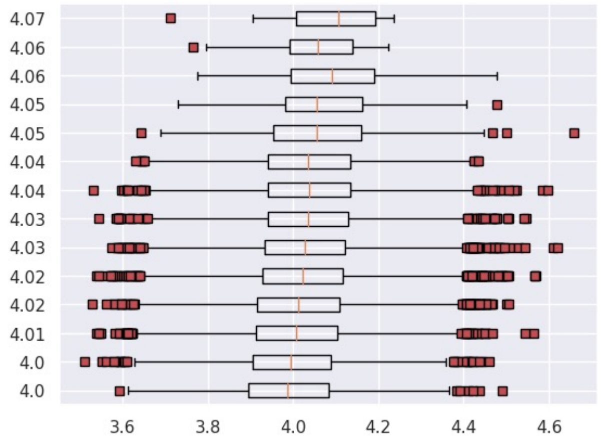


50000 random permutations of 100

Cancellations



Posterior Adjusted Label



Methods

- Large Batch Learning Reduces Regularization

- We are learning from mean of Batches..
  - -> to predict well, prediction needs to be close to Bernoulli -> Overconfidence
  - Model was just doing what it was telling to do

- Mean Label for the batch is Randomly Selected from…
  - Norm(Mean Cancellation Rate, σ * root(n))
  - σ is hyperparameter

- Each elements in batch is randomly selected from training set according to the proportion

Methods

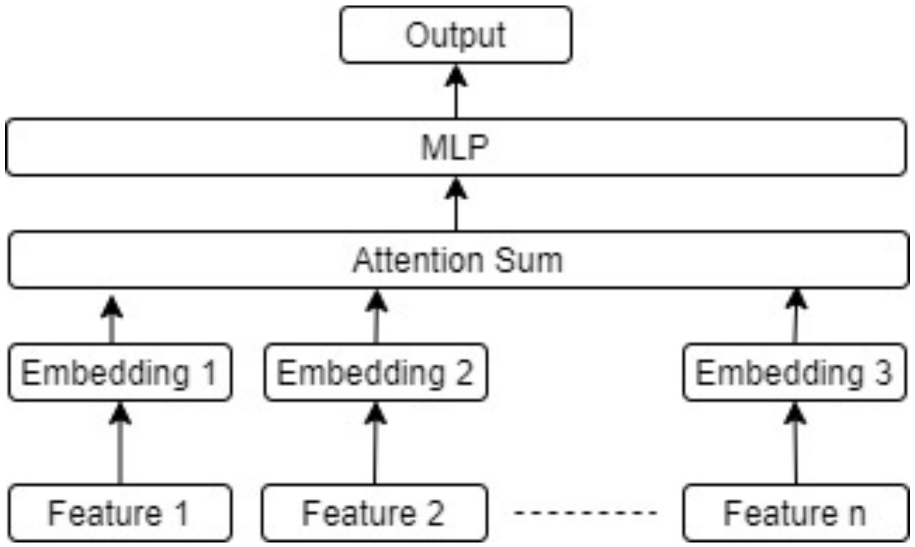
- MixUp
  - One-hot distribution input does not need well defined input boundary

- Training on smoother label
  - Label smoothing is known to help network calibration

- Instead of Mixing two data points from weight selected from Beta Distribution, MixUp multiple data points using weight from Dirichlet Distribution

Methods

- Model
  - Shallow MLP
  - Attention on Feature Embeddings + Shallow MLP
  - DenseNet-121



- Loss
  - BCELoss(MixUp) + BCELoss(Data + Label Smoothing)

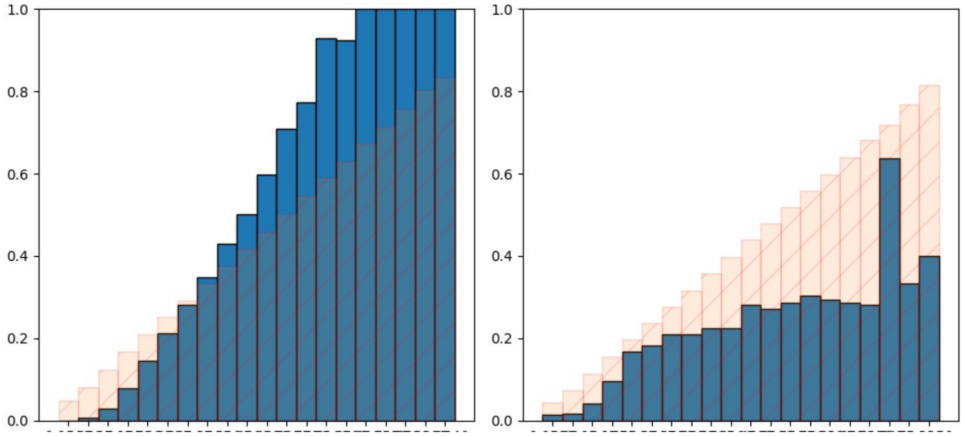
Result

- MLP without feature embedding
  - Overfit, Underconfident on Training set
  - Overconfident on Test set

- MLP with Embedding, Attention
  - Fit on Training set
  - Generally higher prediction on test set
    - 2016~2018 non-show rate: 19.84%
    - 2019 non-show rate: 13.76%

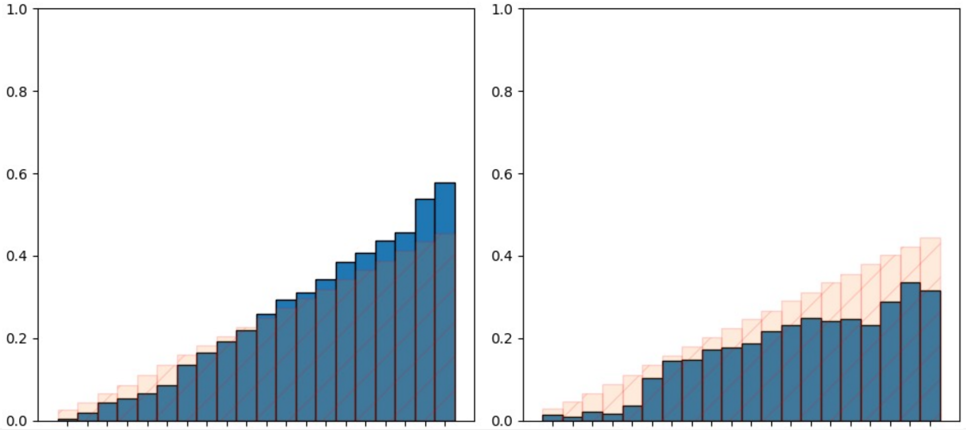
MLP

Training Testing



Embedding Attention Sum + MLP

Training Testing



Result

	Visit mean p	Non-show mean p	Pearson's Correlation	Mean Diff	ECE	entropy	std
naive	0.1984	0.1984	-0.1009	19.31	0.06076	0	0
Mix2,	0.1901	0.2884	0.1444	19.34	0.06602	2.5376	0.1187
Mix2, σ=0.1	0.1764	0.2766	0.2211	16.46	0.05267	2.5191	0.1191
Mix2, σ=0.15	0.1785	0.2762	0.1726	16.91	0.05459	2.4784	0.1192
Mix2, σ=0.2	0.1799	0.2830	0.2118	17.24	0.05644	2.4846	0.1202
Mix4,	0.1810	0.2764	0.1533	17.50	0.05673	2.4269	0.1191
Mix4, σ=0.1	0.1892	0.2888	0.1316	19.41	0.06548	2.4659	0.1196
Mix4, σ=0.15	0.1794	0.2729	0.1795	16.97	0.05489	2.3448	0.1105
Mix4, σ=0.2	0.1861	0.2797	0.1808	18.22	0.06132	2.4276	0.1106

Result

- MixUp
  - Mixing more than 2 input did not help

MixUp	Best Batch Std	Correlation	Mean Diff
2	0.1	0.2211	16.46
3	0.1	0.2193	15.23
4	0.2	0.1808	18.22
5	0.2	0.1456	18.11

- Batch sampling
  - Batch sampling generally improved the result

Thank You

Yunyol Shin
010 8725 2571
yunyol.shin@gmail.com
Bio & Health Informatics Lab