GINA: Group Gender Identification Using Privacy-Sensitive Audio Data

J. Shen, O. Lederman, **J. Cao**, F. Berg, **S. Tang**, A. Pentland

PolyU

mit media lab

UT DALLAS

Outline

I. Background & Motivation
II. Existing works
III. The proposed system
IV. Evaluation

I. Background & Motivation

“ Group Gender Identification Using Privacy-Sensitive Audio Data ”

- Why group gender?
  - Social interaction and group dynamics [1], [2]
  - Foundation of promising research
    (e.g., gender inequality [3] and gender difference [4])

- Why using privacy-sensitive (PS) audio?
  - Spontaneous face-to-face communication in natural settings
  - Ethical issues in collecting the data
  - ✓sampling at 700Hz and averaging amplitude reading every 50 milliseconds [5]

PS audio is to ensure raw audio is not recorded nor can it be reconstructed.

II. Existing works

Vision

Online behavior

Handwriting

Voice

- Voice-based methods
  - Acoustic features caused by physiological differences and phonetic differences
  - Features are extracted from raw audio

- Difficulties caused by PS audio
  - PS audio is too coarse-grained to extract valuable acoustic features
  - Uncertainties caused by natural settings are difficult to address

III. The proposed system

- Problem: Gender identification with PS audio
  - Input: PS audio of a group of people in a meeting
  - Output: gender of each participant

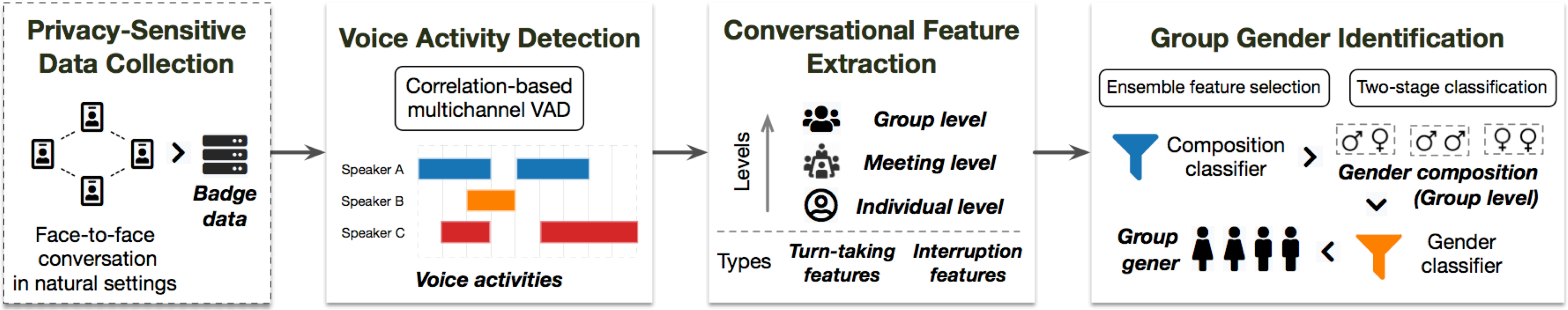
Smart badge for data collection

- Main idea:
  - Conversational behaviors instead of acoustic features

- Challenges
  - C1: Low resolution and unexpected dynamics of PS audio in voice activity detection
  - C2: The instability of conversational behaviors reduces the robustness and effectiveness of gender identification

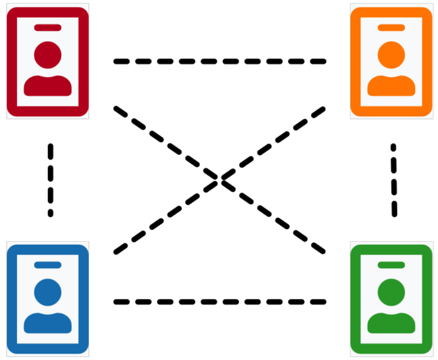
III. The proposed system (cont'd)

- The proposed solutions to the challenges
  - C1: correlation-based multichannel voice activity detection algorithm
  - C2: ensemble feature selection & two-stage classification

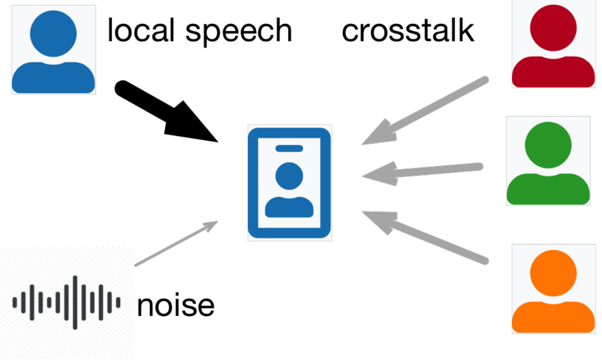


Overview of the proposed system

- Voice activity detection



A meeting with four people

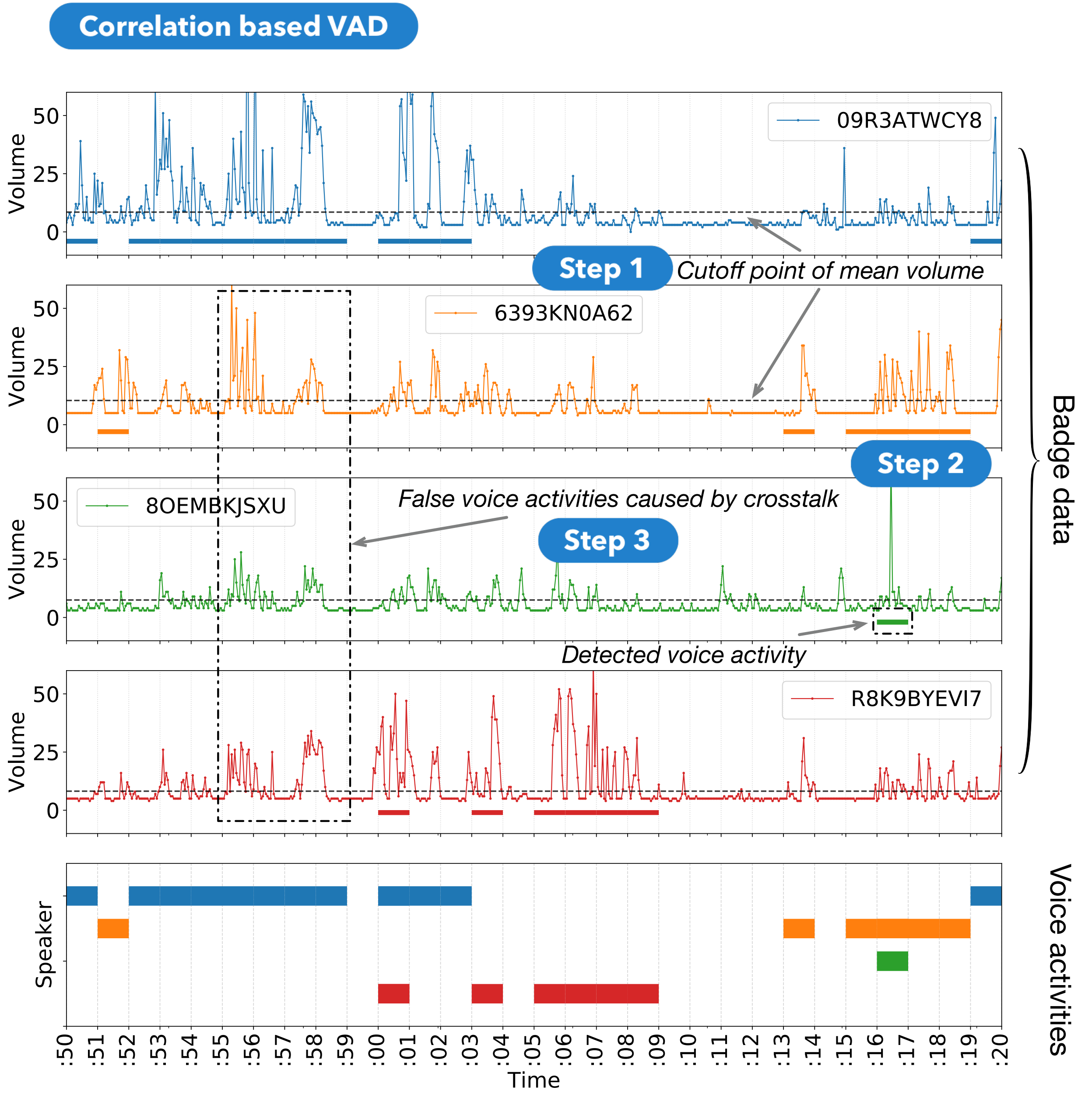


Composition of the badge data

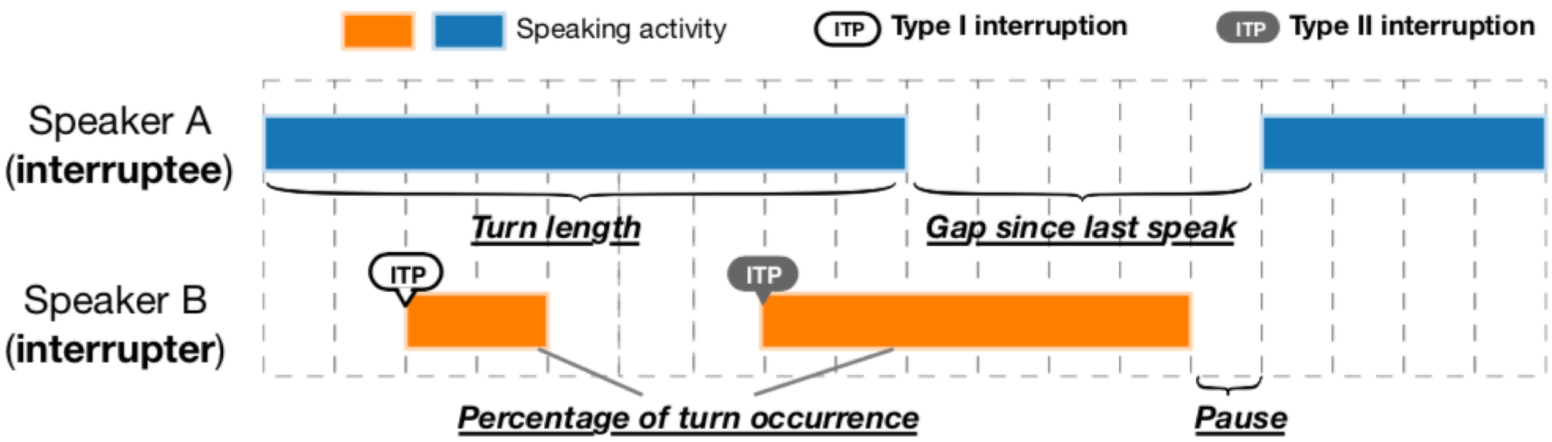
$$\mathbf{S}_i = \underbrace{\boxed{\mathbf{V}_i}}_{\text{Local speech}} + \underbrace{\boxed{\sum_{j \in \mathbf{P}} \phi_{ij} \cdot \mathbf{V}_j}}_{\text{Crosstalk}} + \underbrace{\boxed{\rho_d + \rho_e}}_{\text{Noise}}, j \neq i$$

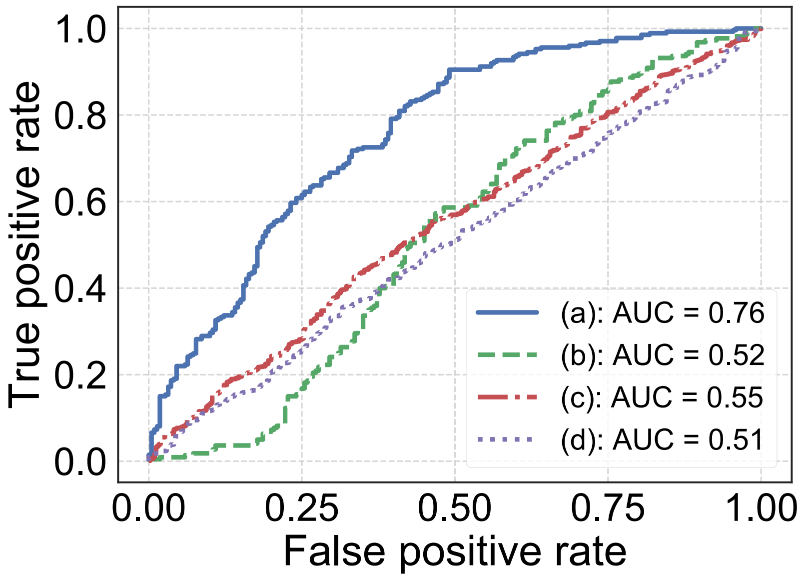
$$\begin{cases} \mathbf{S}_i(k) = \mathbf{V}_i(k) + \rho \approx \mathbf{V}_i(k) \\ \mathbf{S}_j(k) = \phi_{ij} \cdot \mathbf{V}_i(k) + \rho \approx \phi_{ij} \cdot \mathbf{V}_i(k) \end{cases}$$

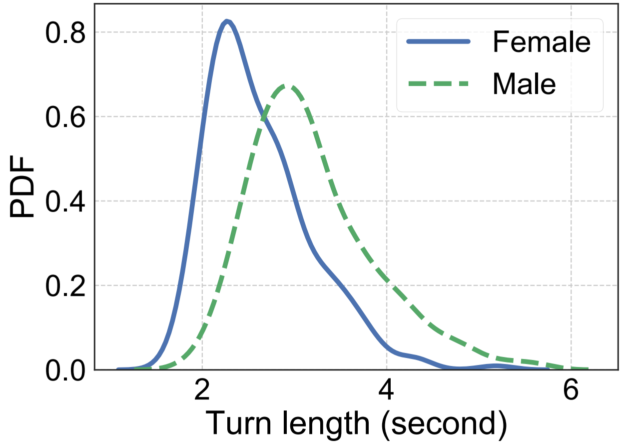
Observation: When only one person speaks, his badge signal is correlated other people's badge signals.



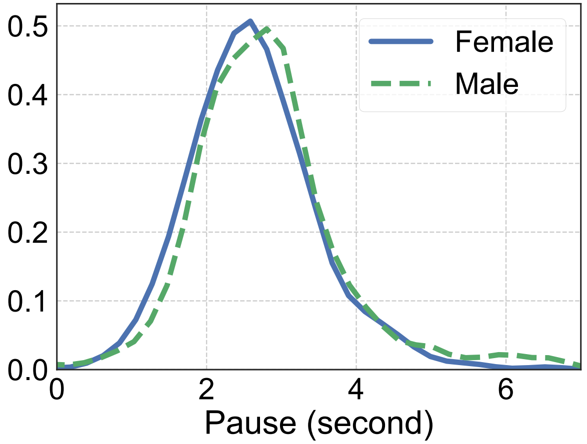
- Conversational feature extraction (turn-taking)



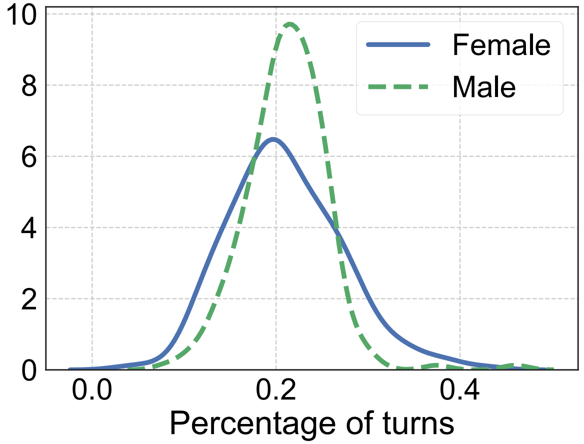




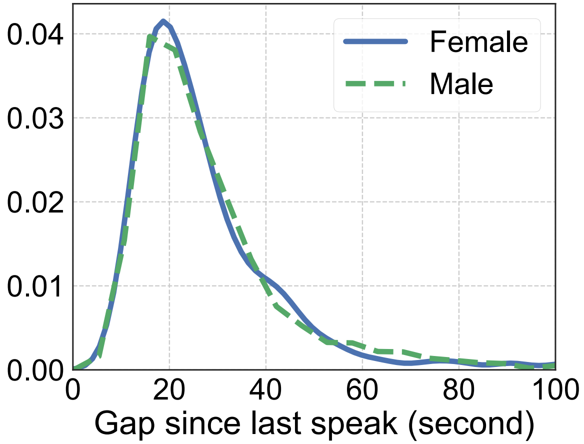
(a)



(b)

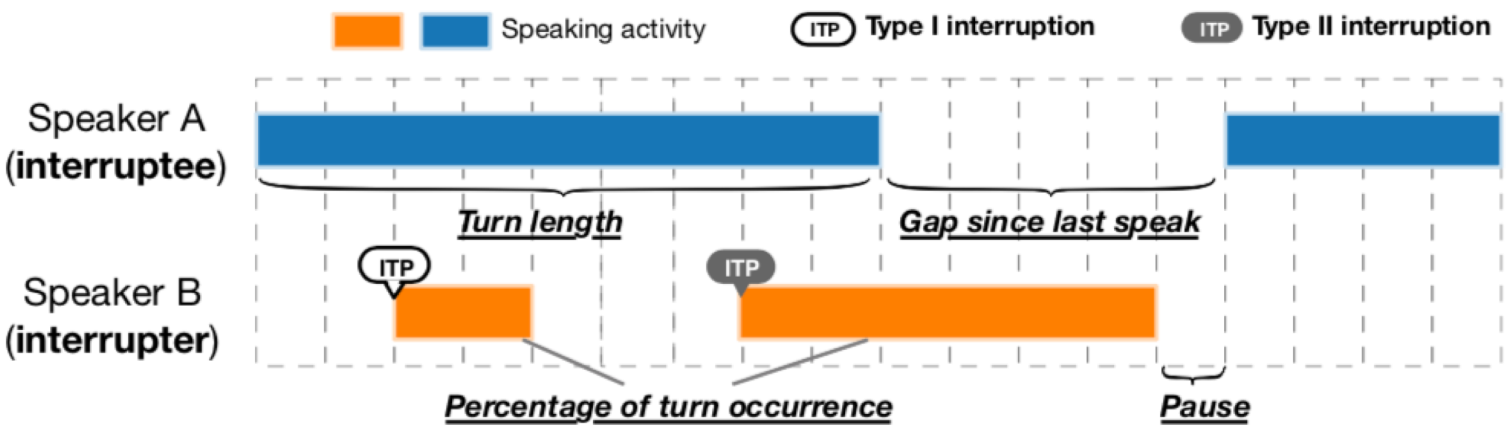


(c)



(d)

- Conversational feature extraction (Interruption)



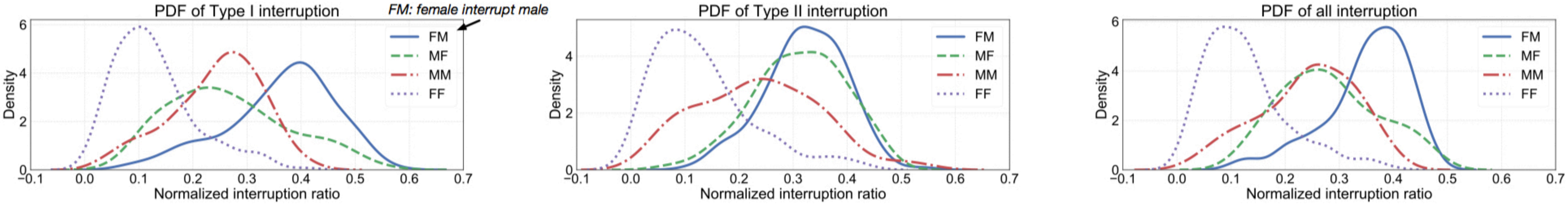
Interruption ratios

$$\begin{array}{c|c} \text{FF} & \text{FM} \\ \hline \text{MF} & \text{MM} \end{array} = \begin{array}{c|c} \frac{\mathbf{I}_{FF}}{\mathbf{I}_F \cdot \mathbf{N}_F} & \frac{\mathbf{I}_{FM}}{\mathbf{I}_F \cdot \mathbf{N}_M} \\ \hline \frac{\mathbf{I}_{MF}}{\mathbf{I}_M \cdot \mathbf{N}_F} & \frac{\mathbf{I}_{MM}}{\mathbf{I}_M \cdot \mathbf{N}_M} \end{array}$$

$\mathbf{I}_{FF}$: Number of FF interruption

$\mathbf{I}_F$: Number interruption started by females

$\mathbf{N}_F$: Number of females in group



	FM	MF	MM	FF
FM		2.7e-16	4.1e-28	8e-53
MF	2.7e-16			1.8e-33
MM	4.1e-28			4.1e-30
FF	8e-53	1.8e-33	4.1e-30	

FM is greater than MF with a p-value 2.7e-16

(a)

	FM	MF	MM	FF
FM			1.5e-15	1.6e-44
MF			6.5e-12	4.6e-41
MM	1.5e-15	6.5e-12		2.6e-18
FF	1.6e-44	4.6e-41	2.6e-18	

(b)

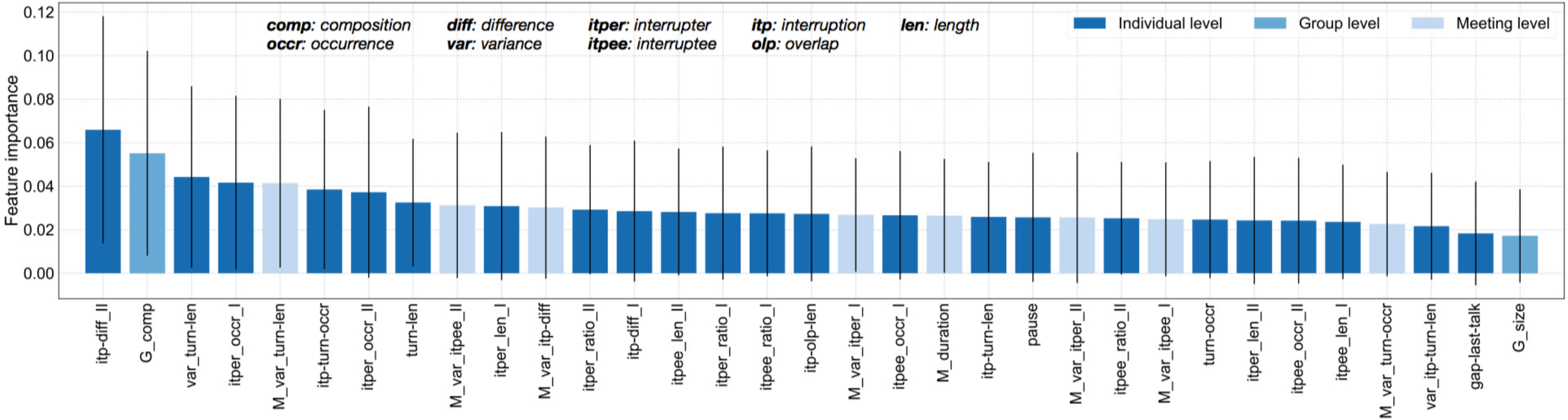
	FM	MF	MM	FF
FM		5.3e-14	6.9e-30	4.5e-55
MF	5.3e-14		0.0003	8e-43
MM	6.9e-30	0.0003		7.9e-29
FF	4.5e-55	8e-43	7.9e-29	

greater

less

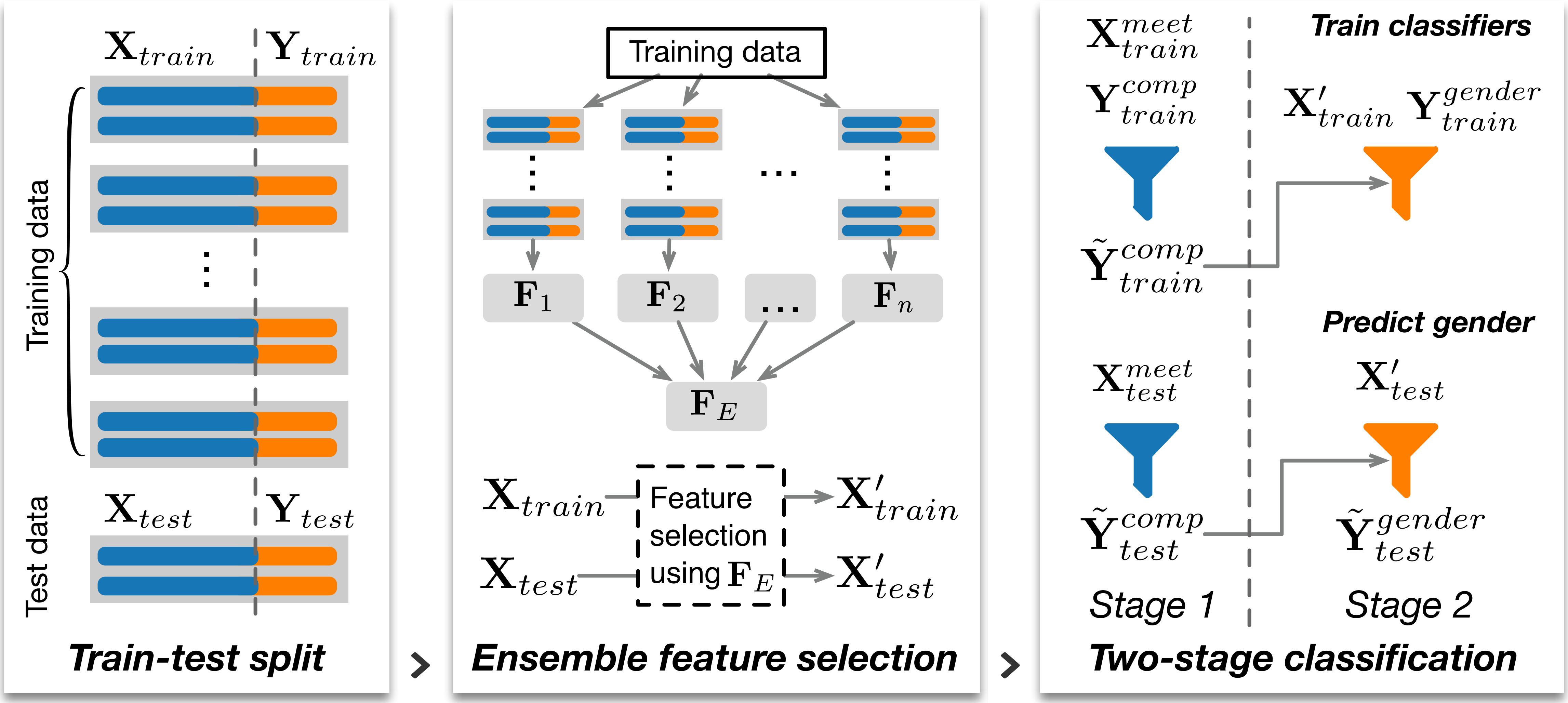
(c)

- Conversational feature extraction (cont'd)



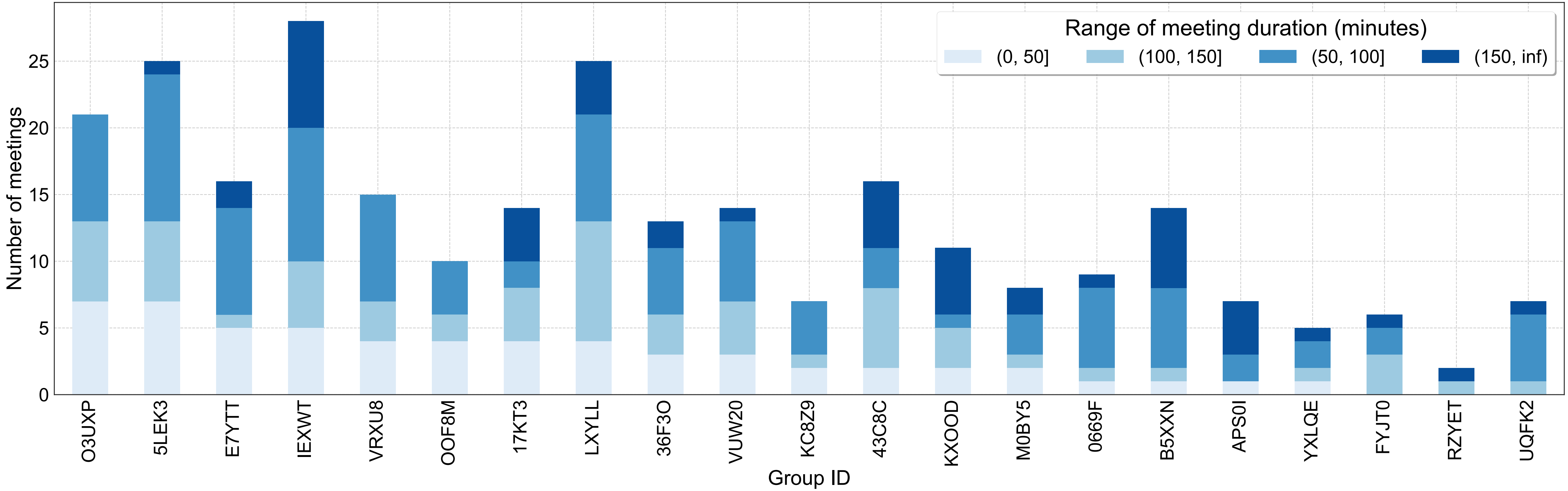
- Hard to find a subset of informative features
- All the features have large variances

- Gender identification



IV. Evaluation

- Dataset
  - **21** study groups, each with **4~5** students (100 in total)
  - **273** effective meetings with a total length of **438.25** hours



IV. Evaluation

Approach	Feature space (in levels)	Feature selection
T-E	Three levels (no composition)	Ensemble feature selection
TC-S	Three levels + composition	Single feature selection
GINA	Three levels + composition	Ensemble feature selection

A *Evaluate group level feature*

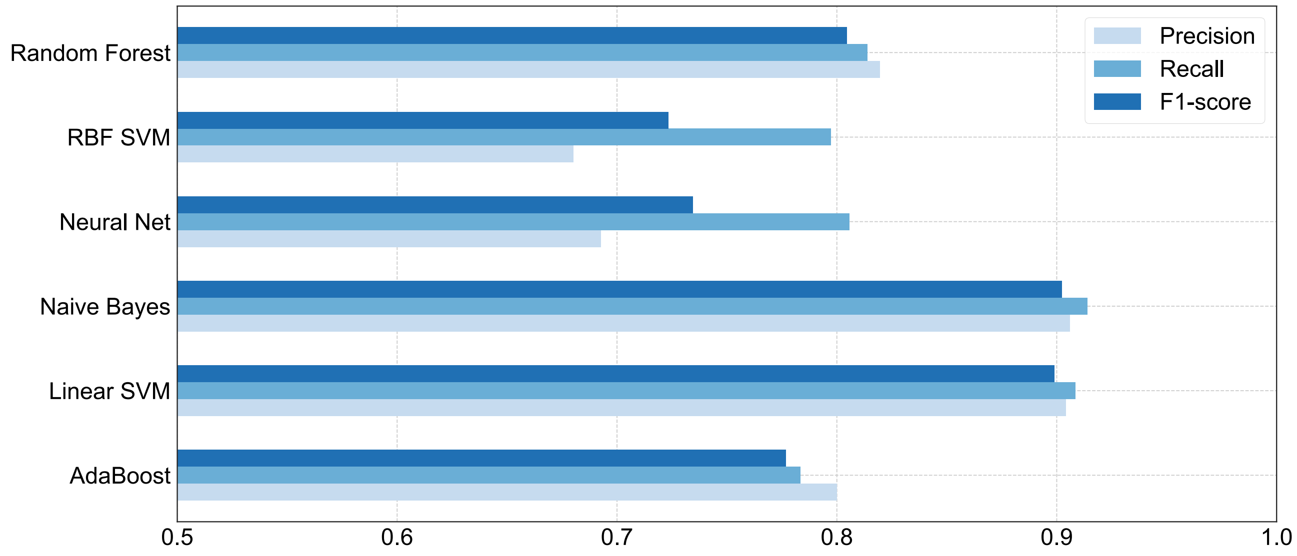
B *Evaluate ensemble feature selection*

Prediction \ Truth	X	$\tilde{X}$
X	tp	fp
$\tilde{X}$	fn	tn

X Target label {female, male}

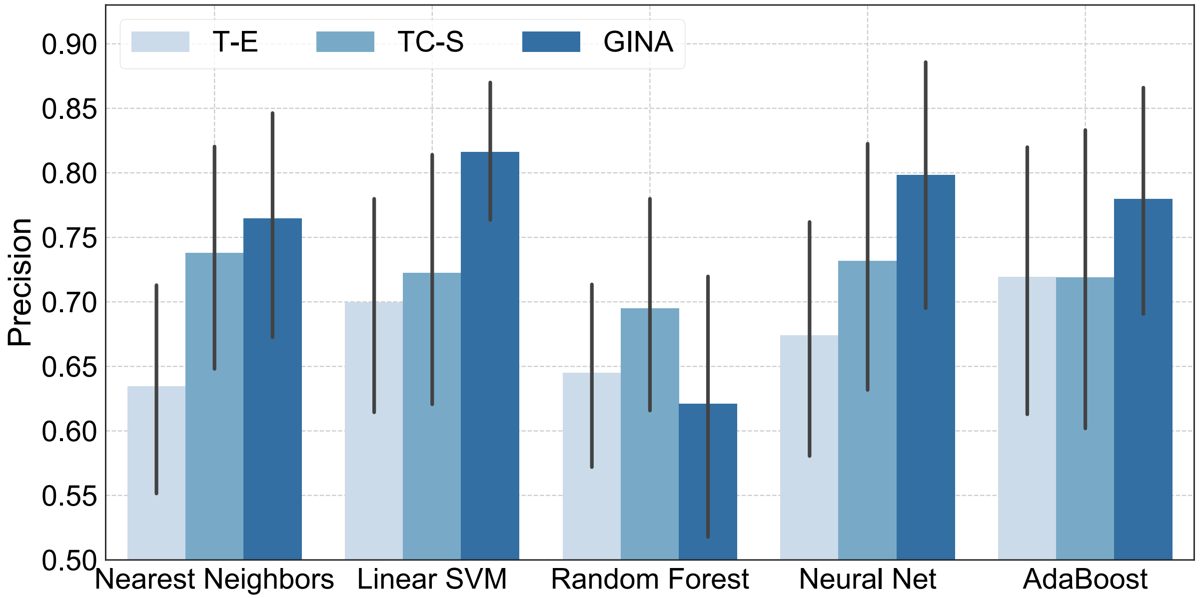
$\tilde{X}$ Non−target label

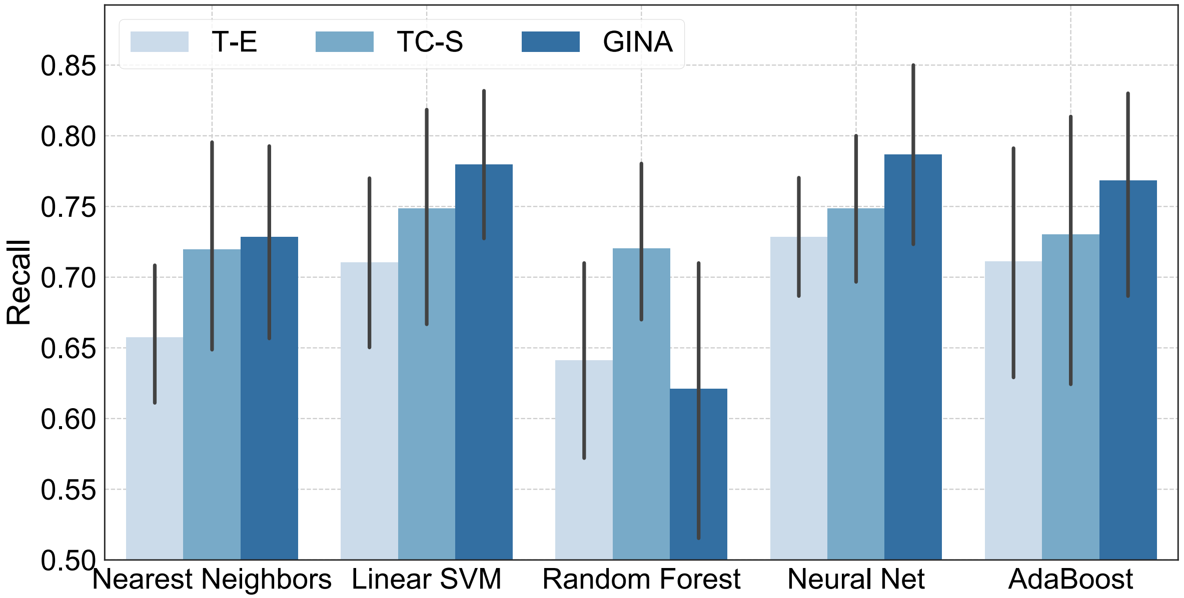
$$\begin{cases} \text{precision}(p) = \dfrac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fp}} \\ \text{recall}(r) = \dfrac{\mathbf{tp}}{\mathbf{tp}+\mathbf{fn}} \\ \text{F1-score} = 2 \cdot \dfrac{p \cdot r}{p+r} \end{cases}$$

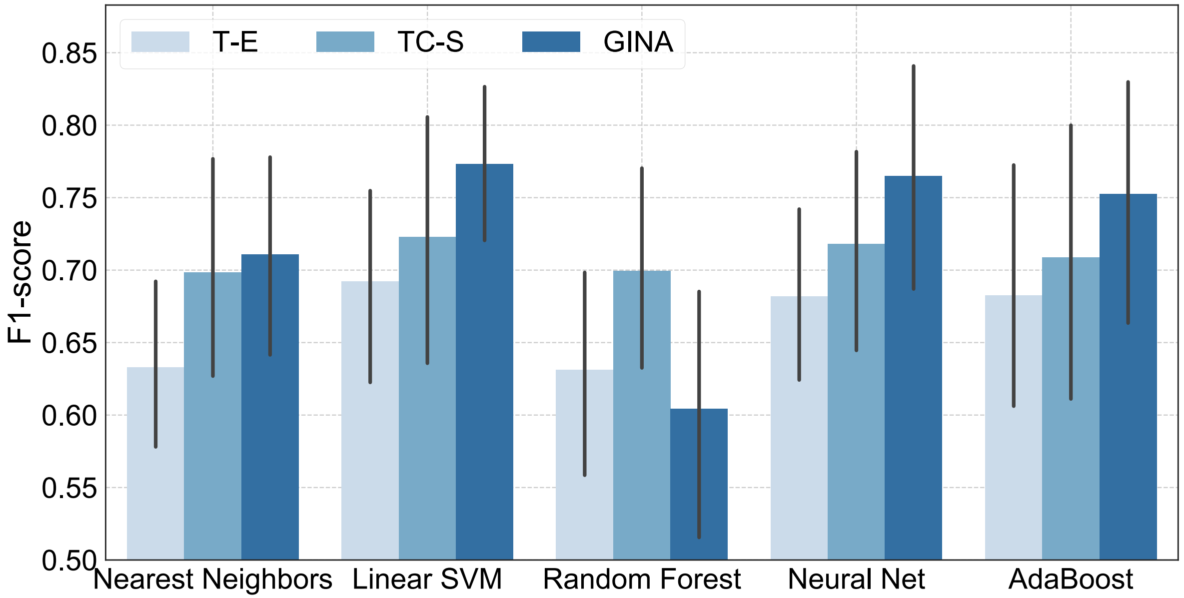


Performance of gender composition detection

IV. Evaluation







References

[1] P. S. Tolbert, M. E. Graham, and A. O. Andrews, “Group gender composition and work group relations: Theories, evidence, and issues,” 1999.

[2] P. Raghubir and A. Valenzuela, “Malefemale dynamics in groups: A field study of the weakest link,” Small Group Research, vol. 41, no. 1, pp. 41–70, 2010.

[3] H. L. Ford, C. Brick, K. Blaufuss, and P. S. Dekens, “Gender inequity in speaking opportunities at the american geophysical union fall meeting,” Nature communications, vol. 9, 2018.

[4] L. Zheng, R. Ning, L. Li, C. Wei, X. Cheng, C. Zhou, and X. Guo, “Gender differences in behavioral and neural responses to unfairness under social pressure,” Scientific reports, vol. 7, no. 1, p. 13498, 2017.

[5] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, “Rhythm: A unified measurement platform for human organizations,” IEEE MultiMedia, vol. 25, no. 1, pp. 26–38, 2018.

Thanks
Questions?