Universiteit Utrecht

Towards Fair and Interpretable Speech-based Depression Severity Modeling

dr. Heysem Kaya

Social and Affective Computing Group, Utrecht University

Social AI Workshop: Social AI for Speech and Conversation

March 28, 2025

Acknowledgements

Gizem Soganciogl u

Albert Ali Salah

Floris van Steijn

Pinar Baki

Outline

- Brief Background on Interpretability
- Interpretability Quest on Bipolar Mania Severity Prediction
- Sequential Concept-bottleneck Models for Interpretability
- Detecting and Mitigating Bias using XAI Methods
- Fairness Perceptions of Mental Health Clinicians

What is Interpretable AI/ML

- No consensus on a universal definition: definitions are domain-specific
- Interpretability: *ability to explain or to present in understandable terms to a human* [A]
  - the degree to which a human can understand the cause of a decision
  - the degree to which a human can consistently predict the outcome of a model
- Explanation: Answer to a *WHY* question
  - relates the feature values of an instance to its model prediction in a humanly understandable way.

[A] Doshi-Velez and Kim, Towards A Rigorous Science of Interpretable Machine Learning, ArXiv 2017

Motivation: Why do we need XAI?

Scientific Understanding
- In search of causal factors / effects

Bias / fairness issues
- Does my model discriminate?

Model debugging and auditing
- Why did my model make this mistake?

Human-AI cooperation / acceptance
- How can I understand / interfere with the model?

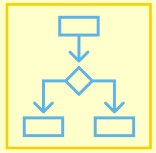
Regulatory compliance
- Does my model satisfy legal requirements? E.g., GDPR*

High-risk applications & regulated industries
- Healthcare, finance / banking, insurance

* https://www.privacy-regulation.eu/en/22.htm

Basic requirements for model interpretability

Intelligible (humanly understandable) features / input



A transparent / simple to understand model (preferably at a glance)

A compact set of predictive features used in the model

Interpretability Request from Medical Domain

- In 2016, we started collaborating with two psychiatrists to answer the following research questions:
  1. Can we find a small set of intelligible speech (acoustic/linguistic) features that can accurately predict mania level in bipolar disorder?

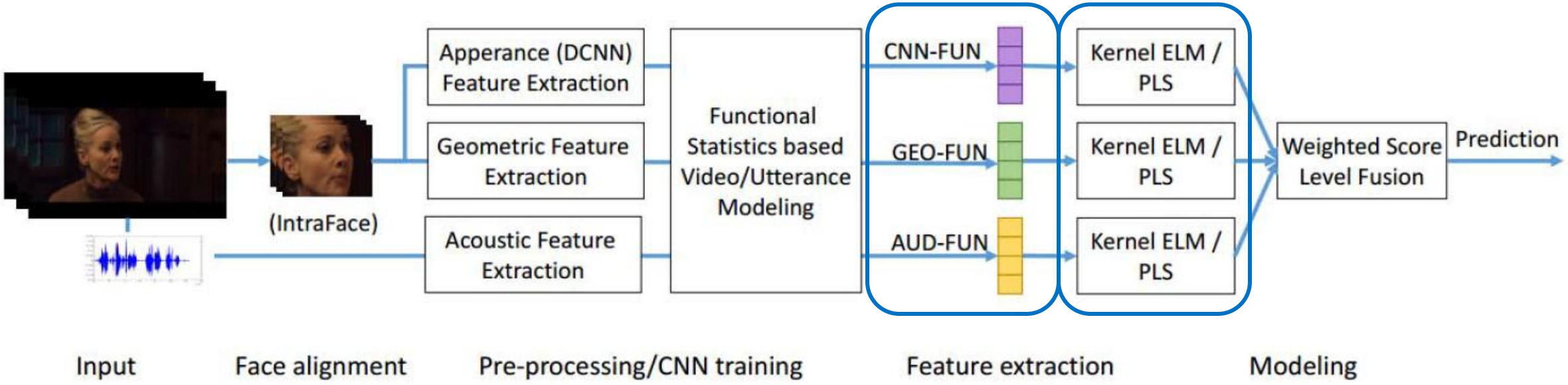
  2. Can we use those features to accurately predict future treatment response?

The Turkish Bipolar Disorder Corpus

- Collecting the Turkish Audiovisual Bipolar Disorder Corpus [A]:
  - 49 patients and 46 healthy controls with similar demographics
  - Annotated for Young Mania Rating Scale (YMRS) [B] for mania severity
  - Multiple sessions at days 0, 3, 7, 14, 28 and 90.

[A] Çiftçi, E., Kaya, H., Güleç, H., & Salah, A. A. (2018). The Turkish Audio-visual Bipolar Disorder Corpus. *ACII Asia.*
[B] Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry.*

How accurate can we get trying to keep the system as simple as possible?



- Best Multimodal UAR (Unweighted Average Recall) Performance [A]:
  - Binary mania detection: ~73% (chance level: 50%)
  - Ternary mania level prediction: ~56% (chance level: 33.3%)

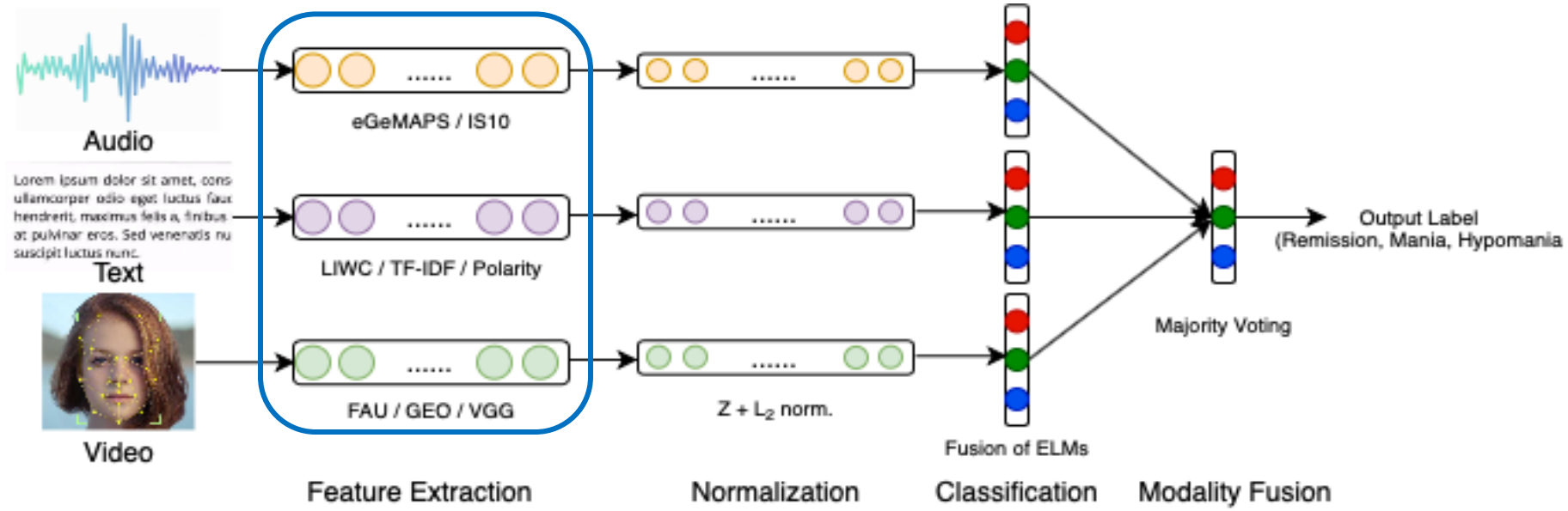
[A] Çiftçi, E., Kaya, H., Güleç, H., & Salah, A. A. (2018). The Turkish Audio-visual Bipolar Disorder Corpus. *ACII Asia.*

How far could others get on Bipolar dataset?

- The dataset was shared in the AVEC 2018 challenge [A] for the ternary mania level prediction task (using only patient data).
- Baseline features and models were provided to the participants.
- Best baseline test performance (57.4% UAR) was obtained with decision fusion of eGEMAPS [B] and Facial Action Unit features.
  - NB: the used features are intelligible to a large extent.
- Challenge result?
- No participant could outperform the baseline test set UAR!

[A] Ringeval, F., et al. (2018). AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*.
[B] Eyben F. et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing.*

Post challenge efforts on Bipolar dataset



We proposed a three-modal system, where in each modality there is at least one (highly) intelligible feature set to allow subsequent interpretability analysis [A].

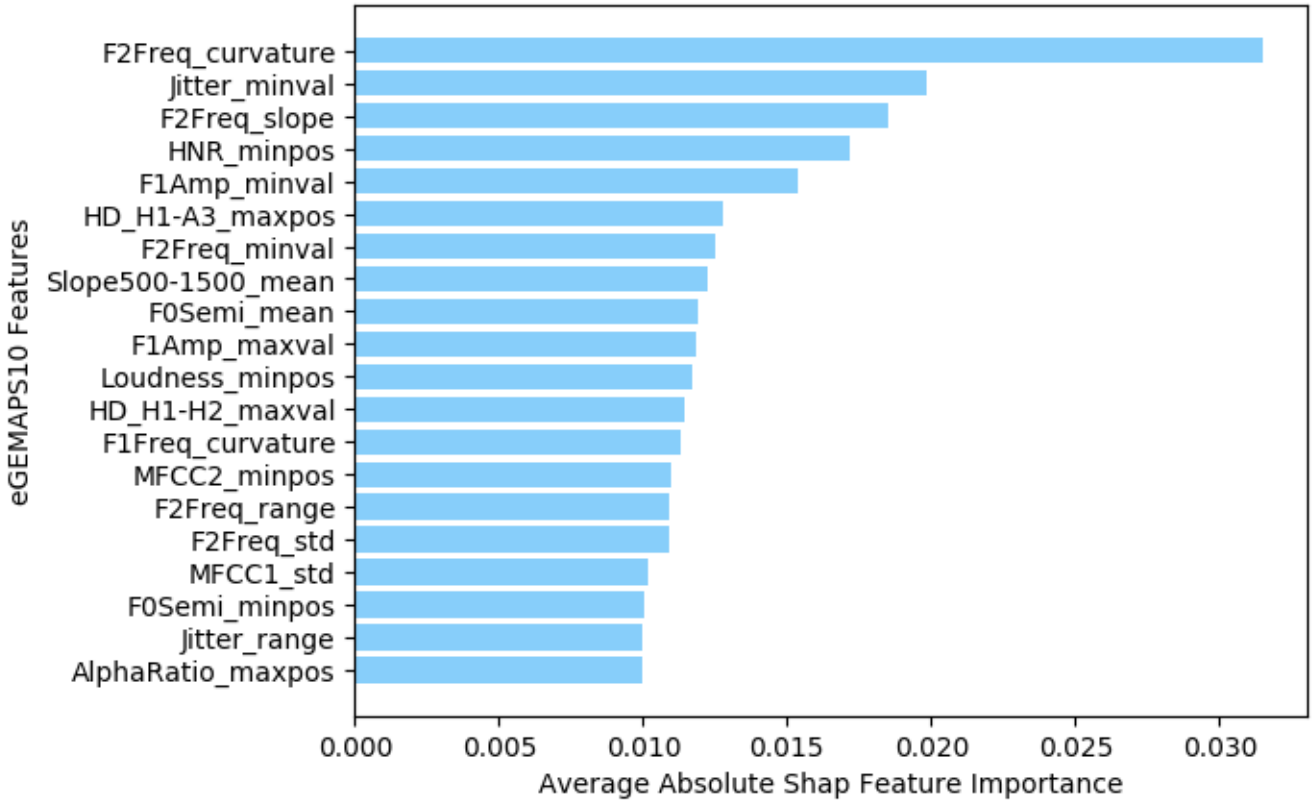
[A] Baki, P., Kaya, H., Çiftçi, E., Güleç, H., & Salah, A. A. (2022). A multimodal approach for mania level prediction in bipolar disorder. *IEEE Transactions on Affective Computing.*

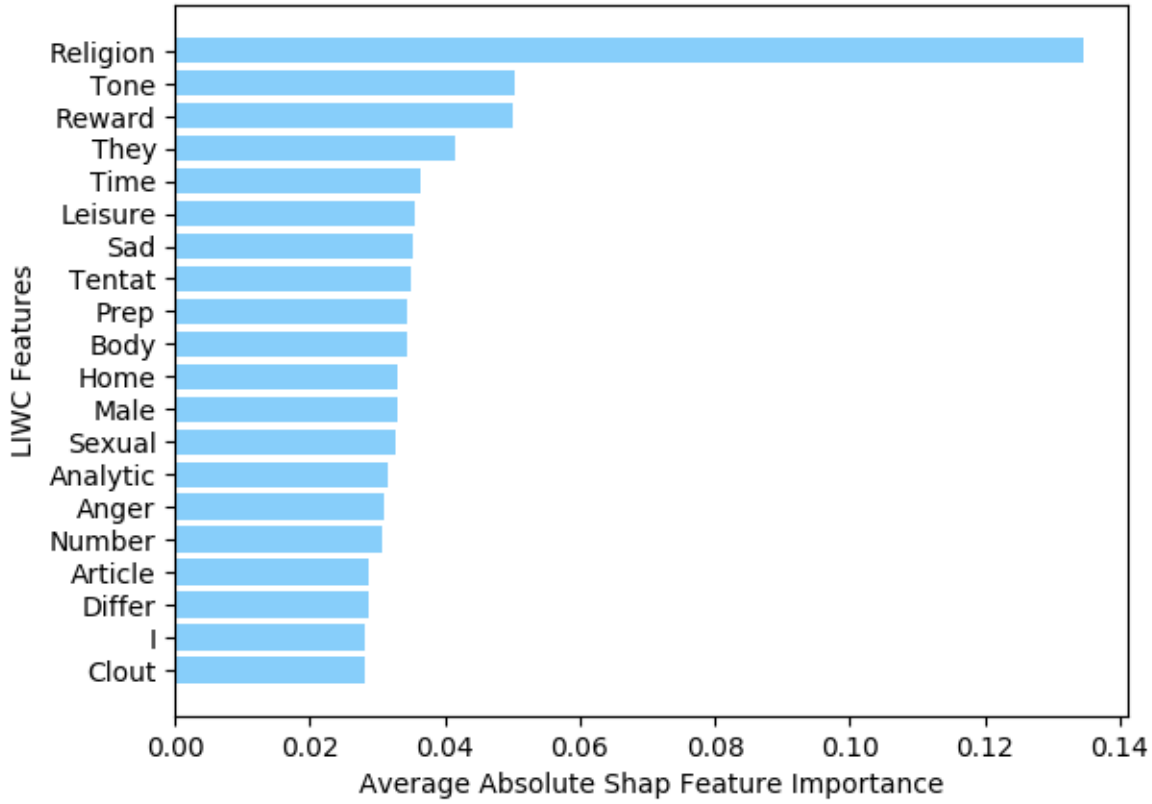
Brief Results on the Challenge Test Set

- Highest unimodal UAR performance: 59.4%
  - eGEMAPS based acoustic model (eGEMAPS10)
- Best multimodal UAR performance: 64.8%
  - Majority voting of modality-specific models trained on **eGEMAPS10, FAU** and **LIWC**
- Advances the SoA UAR but falls below 80% goal!

Baki, P., Kaya, H., Çiftçi, E., Güleç, H., & Salah, A. A. (2022). A multimodal approach for mania level prediction in bipolar disorder. *IEEE Transactions on Affective Computing.*

SHAP-based feature importance analyses





- Most impactful acoustic features include formants, prosody and voice quality features.
- Most impactful LIWC topic is Religion: manic patients that show high value for this feature produce an incoherent discourse intermingled with heavy religious terminology.

Young Mania Rating Scale Item Analysis

YMRS scores are composed of eleven items that assess:

elevated mood (#1), increased motor activity-energy (#2), sexual interest (#3), sleep (#4), irritability (#5), speech rate and amount (#6), language-thought disorder (#7), content (#8), disruptive-aggressive behavior (#9), appearance (#10), and insight (#11).

We analyzed the item activity $a_i = y_i > 0$ prediction performance to get further insights.

4F-CV and Test Set (Last Row) UAR (%) Scores of YMRS Item Activity Prediction Models

Feature/System	YMRS1	YMRS2	YMRS3	YMRS4	YMRS5	YMRS6	YMRS7	YMRS8	YMRS9	YMRS10	YMRS11
eGEMAPS10	66.3	65.4	63.4	**56.9**	65.6	76.4	**72.5**	63.5	**67.4**	68.2	63.6
eGEMAPS	**74.1**	66.2	**69.5**	55.0	68.4	76.4	69.5	**67.9**	61.6	62.3	**64.9**
FAU	68.6	67.4	56.8	56.1	64.8	73.5	60.6	61.6	58.0	69.4	56.6
LIWC	65.7	63.1	61.0	53.3	62.3	69.8	64.8	63.8	56.1	62.5	57.9
MV (eGEMAPS10)	68.5	69.2	61.9	56.2	68.3	**78.1**	70.2	65.8	64.4	**73.8**	61.8
MV (eGEMAPS)	70.1	**69.2**	65.2	55.0	**70.7**	77.8	64.6	66.6	60.6	66.9	59.8
Test Set Performance	64.1	68.6	74.3	62.6	51.7	69.0	67.0	72.0	57.4	46.5	46.5

MV: three modal majority voting.

Young Mania Rating Scale Item Analysis

10. Appearance

0 Appropriate dress and grooming
1 Minimally unkempt
2 Poorly groomed; moderately disheveled; overdressed
3 Disheveled; partly clothed; garish make-up
4 Completely unkempt; decorated; bizarre garb

4F-CV and Test Set (Last Row) UAR (%) Scores of YMRS Item Activity Prediction Models

Feature/System	YMRS1	YMRS2	YMRS3	YMRS4	YMRS5	YMRS6	YMRS7	YMRS8	YMRS9	YMRS10	YMRS11
eGEMAPS10	66.3	65.4	63.4	**56.9**	65.6	76.4	**72.5**	63.5	**67.4**	68.2	63.6
eGEMAPS	**74.1**	66.2	**69.5**	55.0	68.4	76.4	69.5	**67.9**	61.6	62.3	**64.9**
FAU	68.6	67.4	56.8	56.1	64.8	73.5	60.6	61.6	58.0	69.4	56.6
LIWC	65.7	63.1	61.0	53.3	62.3	69.8	64.8	63.8	56.1	62.5	57.9
MV (eGEMAPS10)	68.5	69.2	61.9	56.2	68.3	**78.1**	70.2	65.8	64.4	**73.8**	61.8
MV (eGEMAPS)	70.1	**69.2**	65.2	55.0	**70.7**	77.8	64.6	66.6	60.6	66.9	59.8
Test Set Performance	64.1	68.6	74.3	62.6	51.7	69.0	67.0	72.0	57.4	46.5	46.5

MV: three modal majority voting.

Young Mania Rating Scale Item Analysis

11. Insight

0 Present; admits illness; agrees with need for treatment
1 Possibly ill
2 Admits behavior change, but denies illness
3 Admits possible change in behavior, but denies illness
4 Denies any behavior change

4F-CV and Test Set (Last Row) UAR (%) Scores of YMRS Item Activity Prediction Models

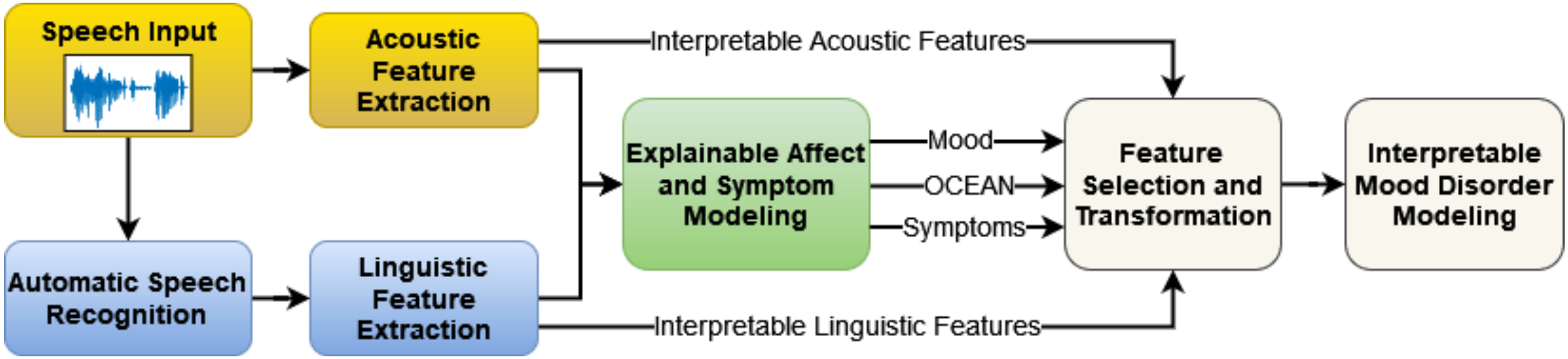
Feature/System	YMRS1	YMRS2	YMRS3	YMRS4	YMRS5	YMRS6	YMRS7	YMRS8	YMRS9	YMRS10	YMRS11
eGEMAPS10	66.3	65.4	63.4	**56.9**	65.6	76.4	**72.5**	63.5	**67.4**	68.2	63.6
eGEMAPS	**74.1**	66.2	**69.5**	55.0	68.4	76.4	69.5	**67.9**	61.6	62.3	**64.9**
FAU	68.6	67.4	56.8	56.1	64.8	73.5	60.6	61.6	58.0	69.4	56.6
LIWC	65.7	63.1	61.0	53.3	62.3	69.8	64.8	63.8	56.1	62.5	57.9
MV (eGEMAPS10)	68.5	69.2	61.9	56.2	68.3	**78.1**	70.2	65.8	64.4	**73.8**	61.8
MV (eGEMAPS)	70.1	**69.2**	65.2	55.0	**70.7**	77.8	64.6	66.6	60.6	66.9	59.8
Test Set Performance	64.1	68.6	74.3	62.6	51.7	69.0	67.0	72.0	57.4	46.5	46.5

MV: three modal majority voting.

Bipolar Disorder: Closing

- RQ: Could we find a small set of intelligible speech features that can accurately predict mania level in bipolar disorder?
- Answer: Not really
- Q: What experiences did we learn for future research?
- Challenges in data collection for mental healthcare
- Increasing need to focus on interpretability
- **Breaking down the problem into recognition of more observable, verifiable cues (e.g., emotion, symptoms) is worth investigation**
  - **Using deep and accurate but explainable models for certain sub-tasks.**

A proposal for Interpretable Speech-based Mood Disorder Prediction

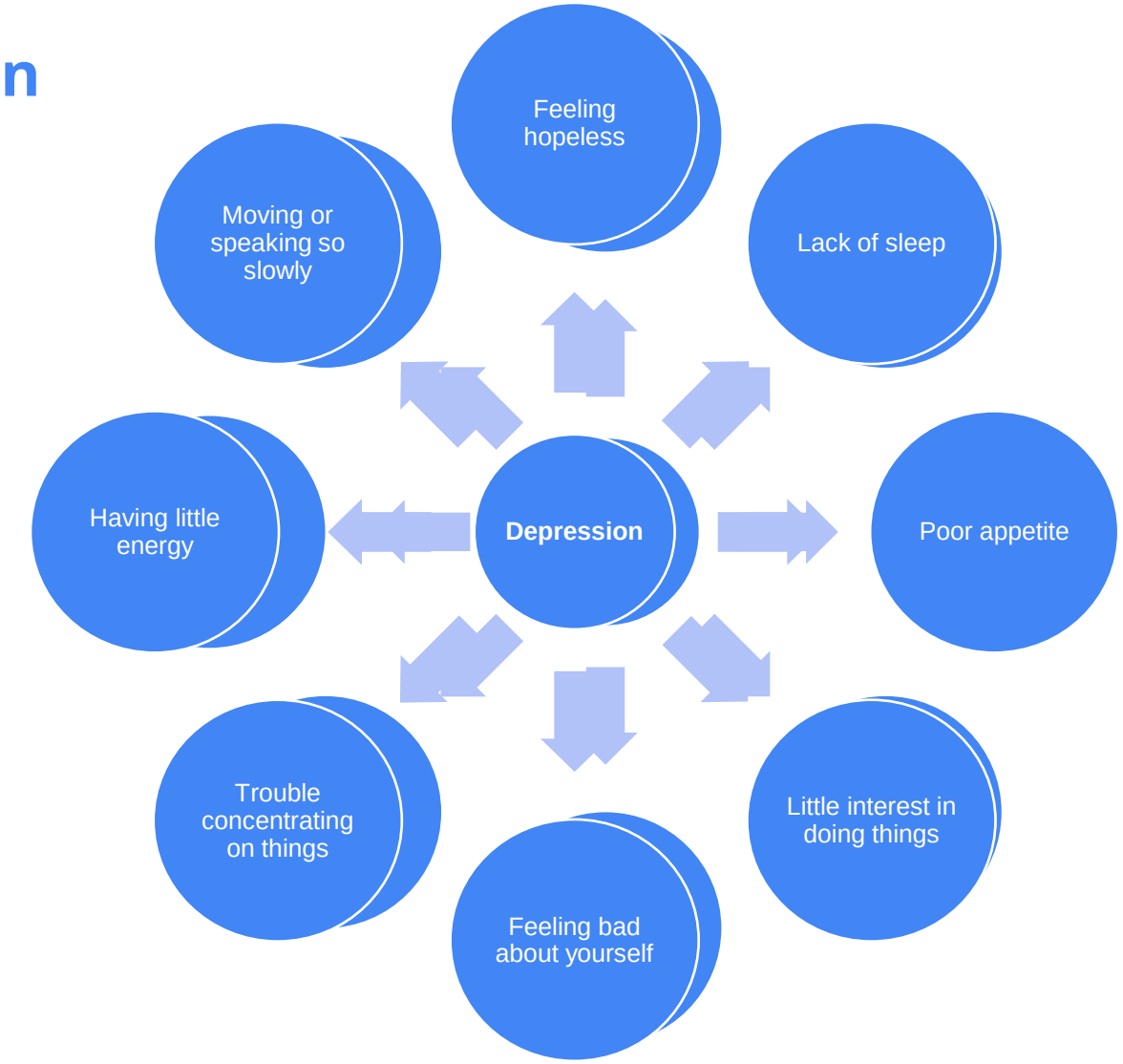


Kaya, H. 'InterpretME: Actionable and Interpretable Paralinguistic Modeling for Mood Disorders', Project proposal

Symptom-based Depression Severity Prediction

Another mood disorder: depression

Depression is common in society.



van Steijn, F., Sogancioglu, G., & Kaya, H. (2022). Text-based interpretable depression severity modeling via symptom predictions. ICMI.

feeling hopeless

moving or speaking so slowly

lack of sleep

Break down the problem into **observable** and **verifiable** components

poor appetite

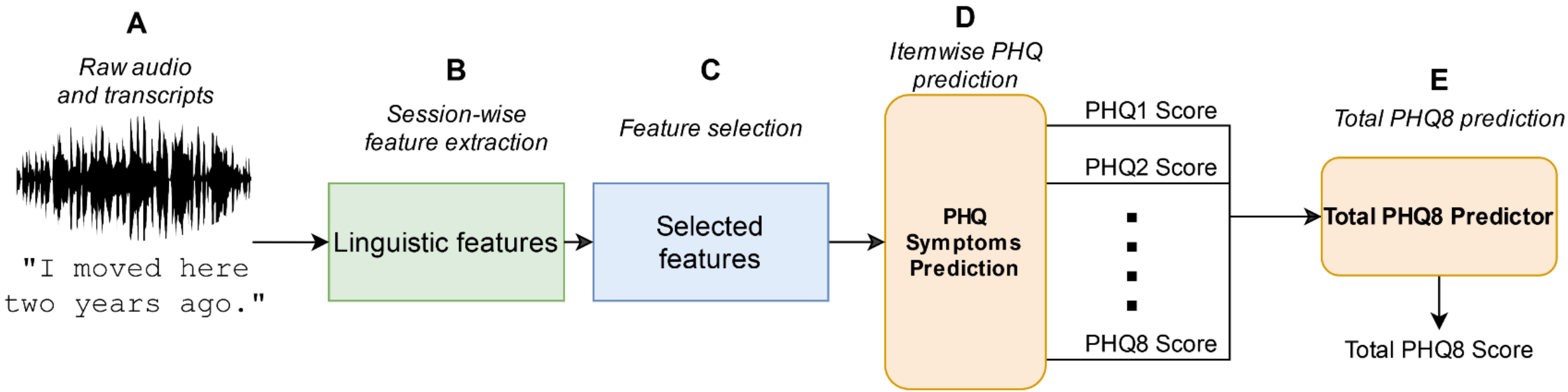
having little energy

little interest in doing things

feeling bad about yourself

trouble concentrating on things

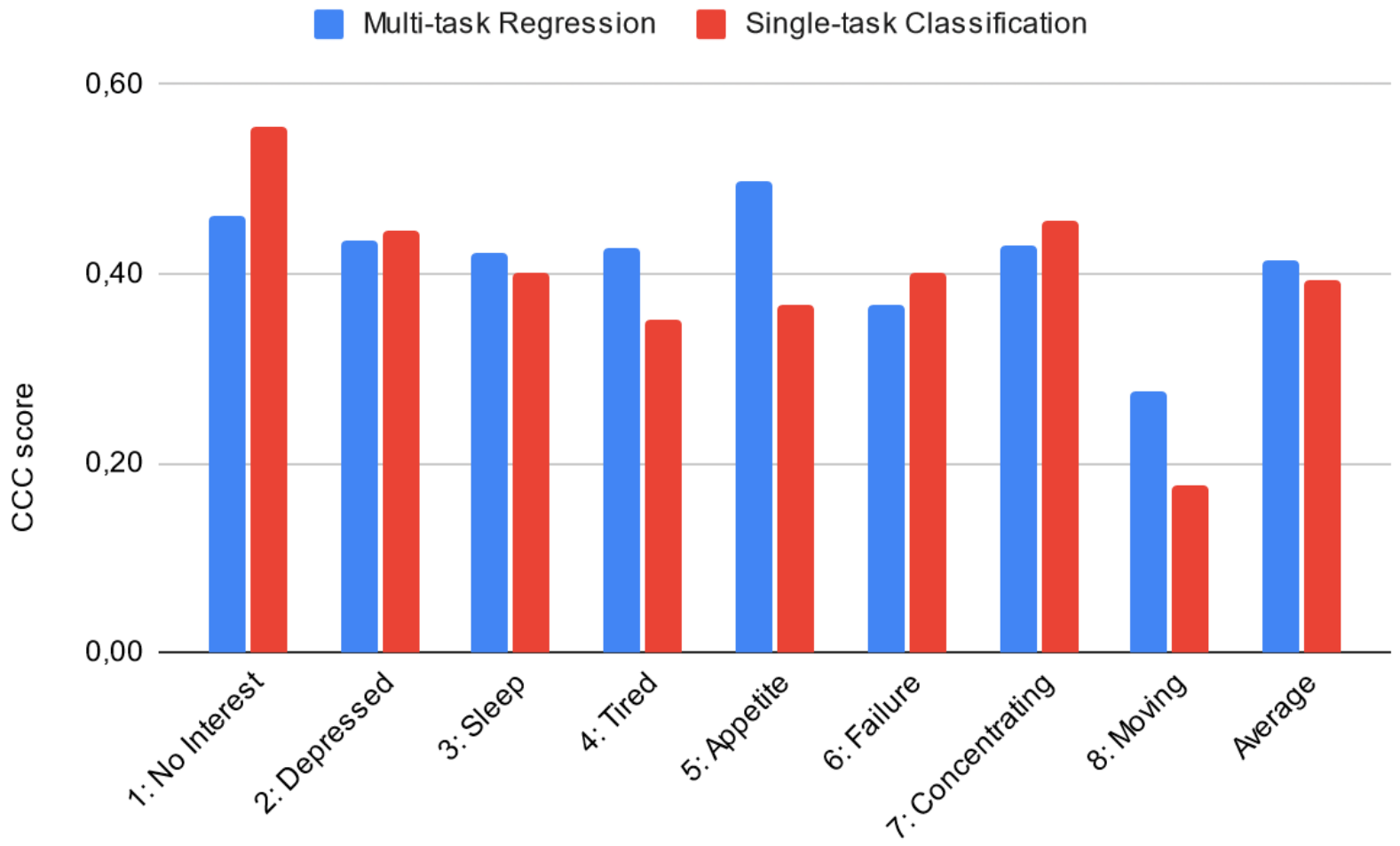
Symptom-based Depression Severity Prediction



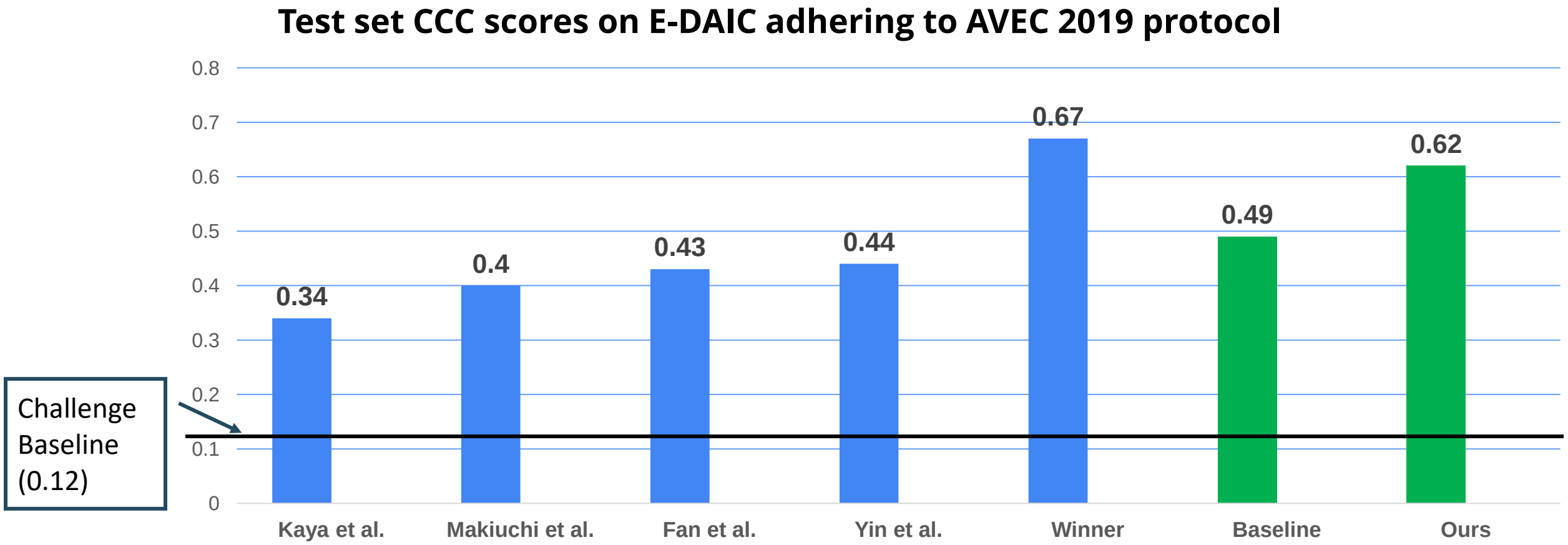
Linguistic feature sets: Sentence-BERT [A], LIWC and handcrafted features (e.g., speech rate, repetition rate).

van Steijn, F., Sogancioglu, G., & Kaya, H. (2022). Text-based interpretable depression severity modeling via symptom predictions. ICMI.
[A] Reimers N., Gurevych I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proc. EMNLP.

Symptom Prediction Performance on E-DAIC corpus



van Steijn, F., Sogancioglu, G., & Kaya, H. (2022). Text-based interpretable depression severity modeling via symptom predictions. ICMI.



Our **Baseline**: direct prediction of total PHQ8 score using the same set of features and regressor.

Summary

Symptom-based Depression Severity Modeling

- **Contribution:** competitively performing interpretable depression severity prediction models using human understandable features.
- **Limitation:** Providing interpretability to only the second-tier models.

Predicting and Using Big-Five Traits (OCEAN)

Trait	Low Scorers	High Scorers
Openness to Experience	Down-to-earth Uncreative Conventional Uncurious	Imaginative Creative Original Curious
Conscientiousness	Negligent Lazy Disorganized Late	Conscientious Hard-working Well-organized Punctual
Extroversion	Loner Quiet Passive Reserved	Joiner Talkative Affective Affectionate
Agreeableness	Suspicious Critical Ruthless Irritable	Trusting Lenient Soft-hearted Good-natured
(Non-)Neuroticism	Worried Temperamental Self-conscious Emotional	Calm Even-tempered Comfortable Unemotional

Literature: Big-Five Traits in Mood Disorders

- Clinical association of personality traits with mood disorders:
  - All disorders have a configuration of low Conscientiousness and high Neuroticism.
- Major Depression Disorder is found to have the strongest correlation with the Neuroticism factor.
- An analysis by Malouff et al. showed a common five-factor configuration of
  - high Neuroticism,
  - low Conscientiousness,
  - low Agreeableness, and
  - low Extraversion for almost all disorders.

Malouff, J. M. et al. (2005). The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of psychopathology and behavioral assessment.*

Explainable Job Interview Recommendation

- In 2017, ChaLearn Looking at People Workshop at CVPR:

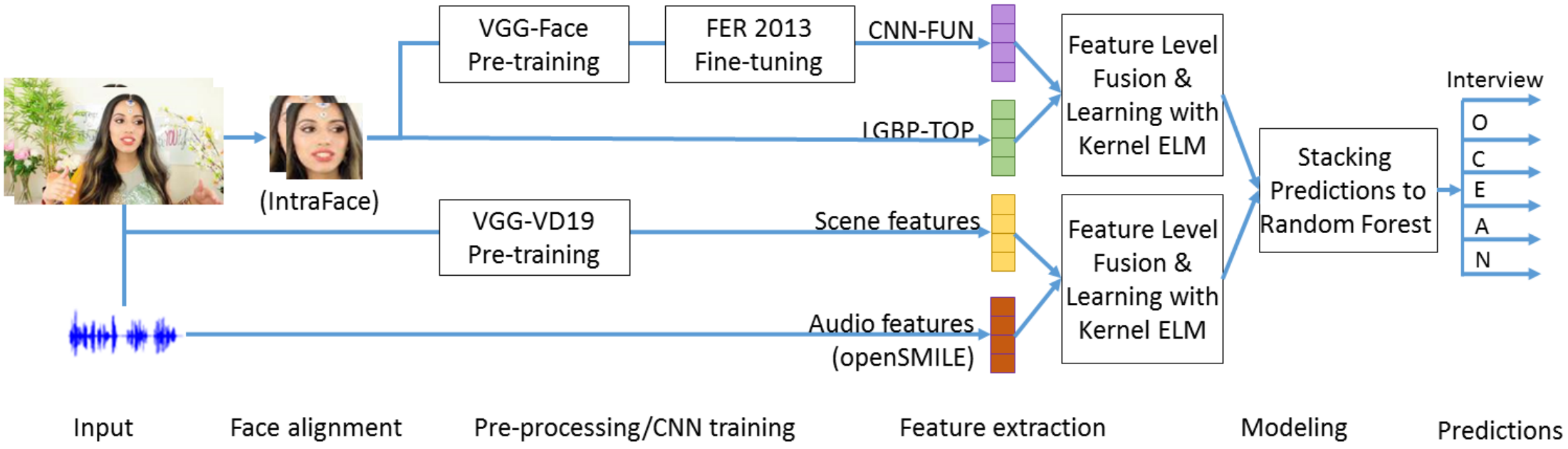
  ‘Explainable Job Interview Recommendation Competition’

- Tasks: predicting OCEAN impressions and interview recommendation
  - Quantitative challenge: predicting interview invitation score given a short video.
  - Qualitative challenge: providing interpretation for the model & decision

https://chalearnlap.cvc.uab.cat/dataset/24/description/

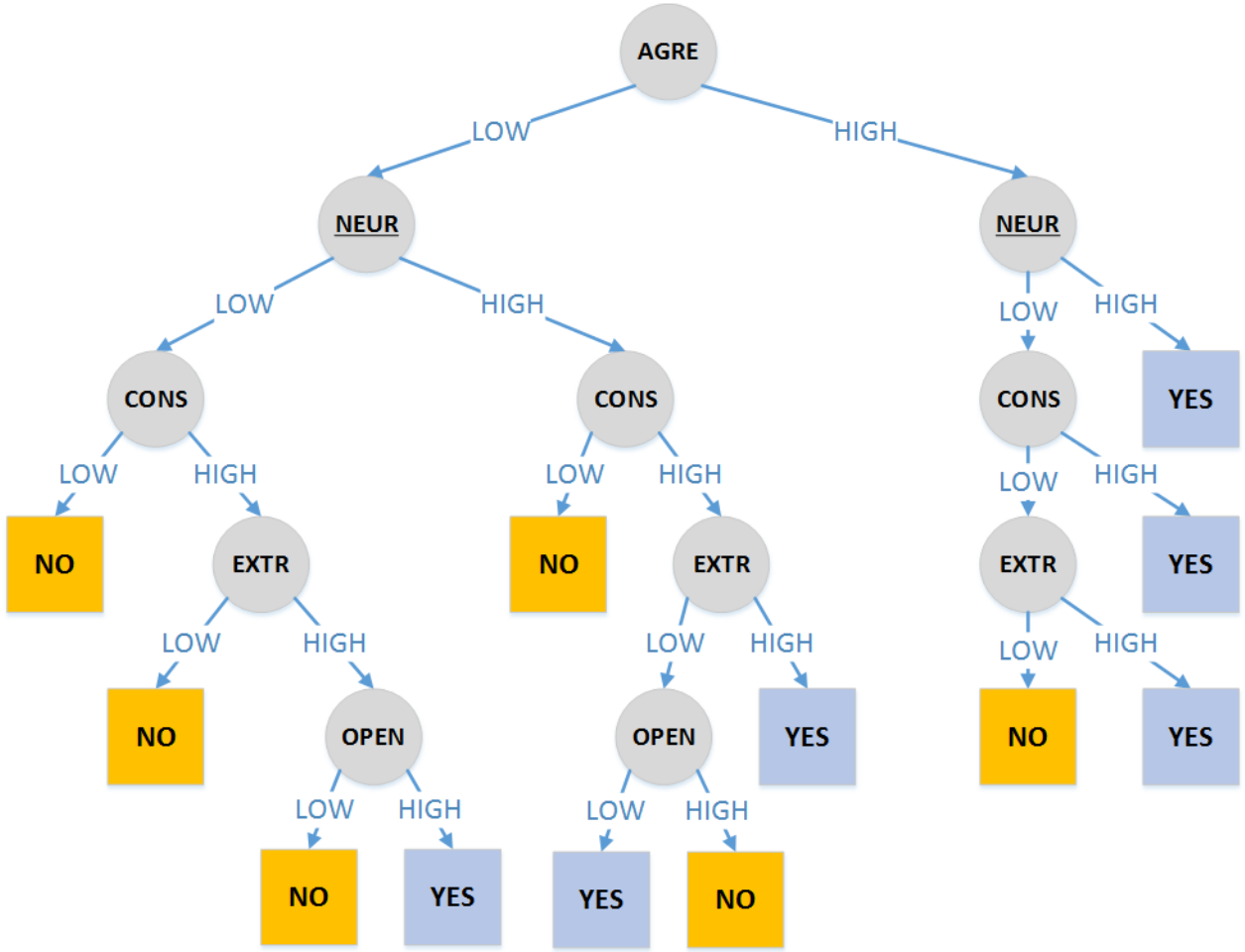
Predicting the Personality Impressions

- Contribution to CVPRW'17 ChaLearn Explainable Job Interview Recommendation Competition [A]



[A] Kaya et al., Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. *CVPRW* 2017.

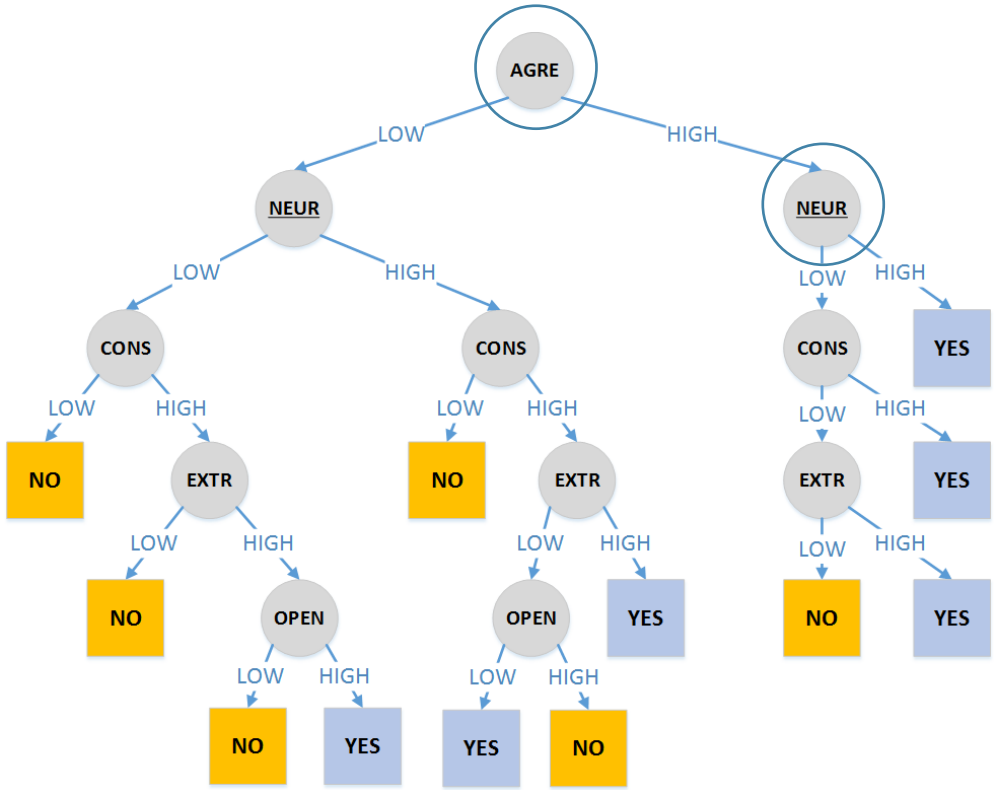
Interview Invitation Explanation Model



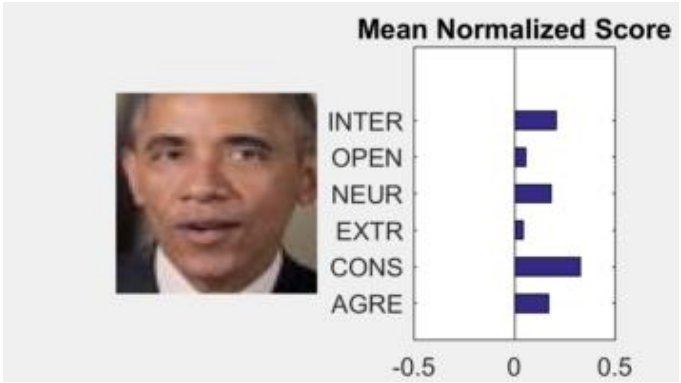
Kaya et al., Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. *CVPRW* 2017.

Providing model & instance interpretability

- Model interpretation → illustration of the tree
- Instance explanation → conjuction of the nodes from root to leaf



Visual Explanation



Automatic Verbal Explanation

This gentleman is invited for an interview due to his **high apparent agreeableness and non-neuroticism** impression.

Evaluation Measures for Qualitative Challenge

- **Clarity**: Is the text understandable / written in proper English?
- **Explainability**: Does the text provide relevant explanations to the hiring decision made?
- **Soundness**: Are the explanations rational and, in particular, do they seem scientific and/or related to behavioral cues commonly used in psychology.
- **Model interpretability**: Are the explanations useful to understand the functioning of the predictive model?
- **Creativity:** How original / creative are the explanations?

Results?

The Bias Problem

Pearson Correlations Among Traits and Personality Impressions

Correlation	Gender	Ethnicity		
Dimension	Female	Asian	Caucasian	Afro-American
Agreeableness	-0.023	-0.002	**0.061****	**-0.068****
Conscientiousness	**0.081****	0.018	**0.056****	**-0.074****
Extroversion	**0.207****	**0.039***	**0.039***	**-0.068****
Neuroticism	**0.054***	-0.002	**0.047***	**-0.053****
Openness	**0.169****	0.010	**0.083****	**-0.100****
Interview	**0.069****	0.015	**0.052***	**-0.068****

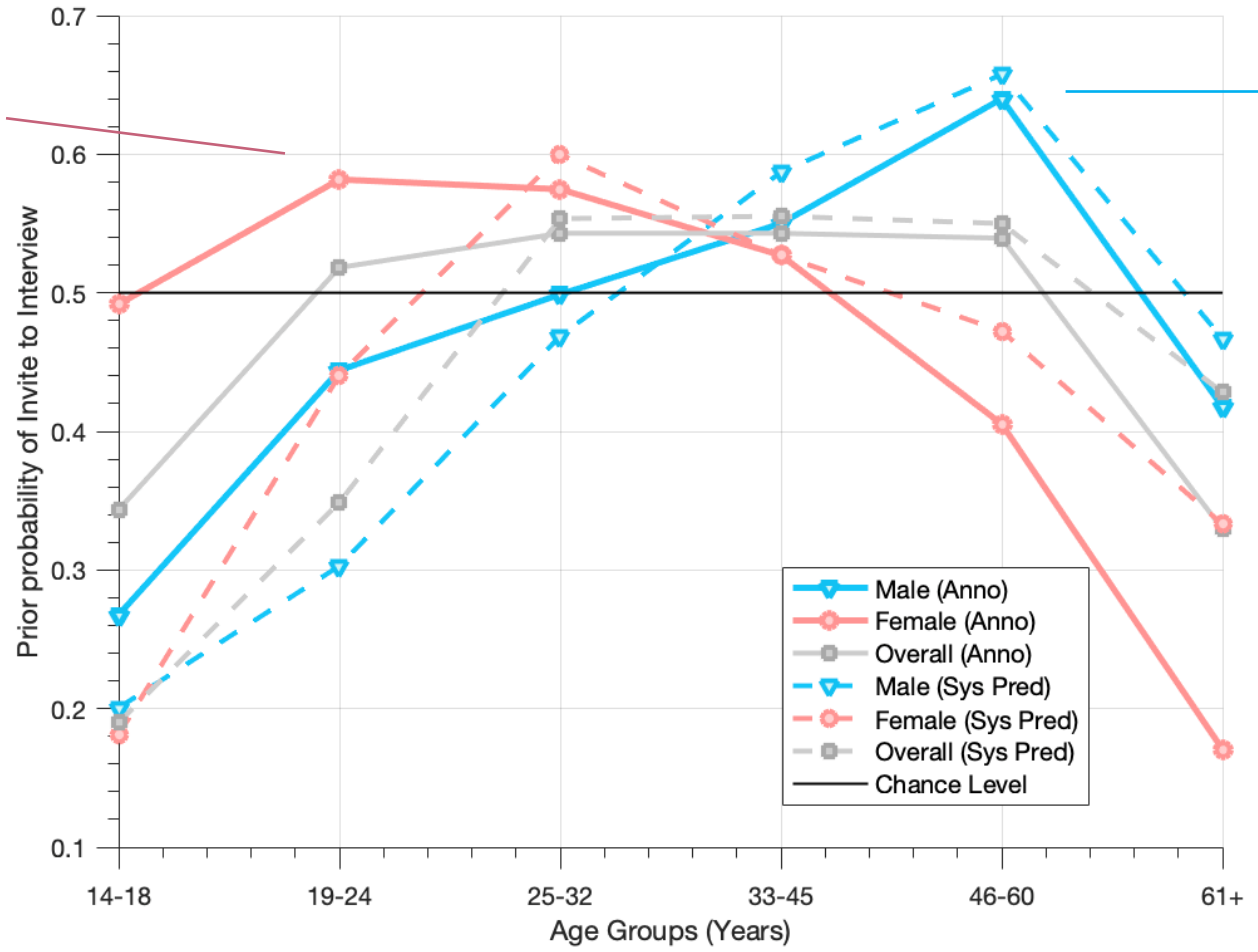
Prior Probabilities of «Invite to Interview»

	Male	Female	Asian	Caucasian	Afro-American
mean scores	0.539	0.589	0.515	0.507	0.475
p(invite | trait)	0.495	0.560	0.562	0.539	0.444

Escalante, Kaya, Salah et al., Modeling, recognizing, and explaining apparent personality from videos*, Transactions on Affective Computing*, 2022

Accurate, explainable but clearly biased!

Younger females are preferred



Older males are preferred

The AI algorithm just learns human bias!

Escalante, H.J., Kaya, H., Salah, A.A., Escalera, S., Gucluturk, Y., Guclu, U., Baró, X., Guyon, I., Junior, J.J., Madadi, M. and Ayache, S., Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. *IEEE Trans. Affective Computing, 2022*.

Summary

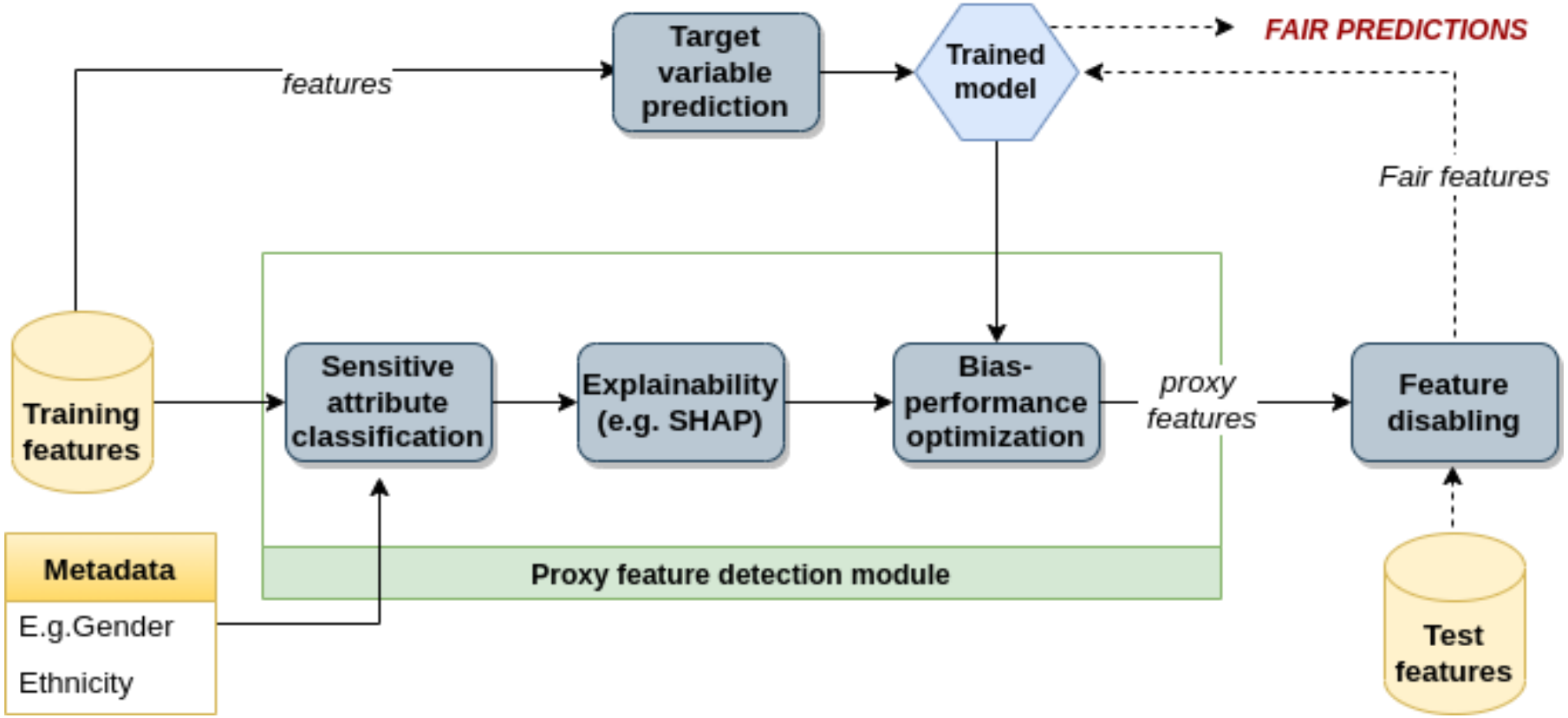
Explainable Job Interview Recommendation

- **Contribution:** More observable concepts such as personality impressions help in accurate and interpretable modeling of more complex tasks.
- **Limitations:** The personality impression predictions are themselves
  - biased for ethnicity and age-gender attributes
  - not explainable in the proposed work
- **Important to note:**
  - We advice against the use of such systems in workplace!
  - EU AI Law prohibits such uses of AI in workplace environment.

Detecting and Mitigating Bias

ProxyMute*

A **bias mitigation algorithm** that uses explainability methods to eliminate sensitive information in feature representation.



*Sogancioglu, G., Kaya, H., & Salah, A. A. (2023). Using Explainability for Bias Mitigation: A Case Study for Fair Recruitment Assessment. In *Proc. ICMI* .

Feature disabling algorithm

Assumptions : feature independence & feature linearity

- a *continuous* feature: the training set **mean** value.
- an *ordinal* feature: the training set **median** value.
- a *categorical* or *binary* feature, the training set **mode** value during the inference time

Sogancioglu, G., Kaya, H., & Salah, A. A. (2023). Using Explainability for Bias Mitigation: A Case Study for Fair Recruitment Assessment. In *Proc. ICMI* .

Recruitment assessment

Dataset: ChaLearn LAP-FI & Interview

Features: Acoustic features + visual features from [A] (26853D)

Model: XGBoost (tree ensemble)

Demographics: gender, race, age

Bias types:

- Feature bias (F)
- Labelling bias (L)
- Sampling bias (S)

Fairness Measure LAP-FI Dataset

Equal Accuracy (EA): the difference in the performance measure between sensitive groups.

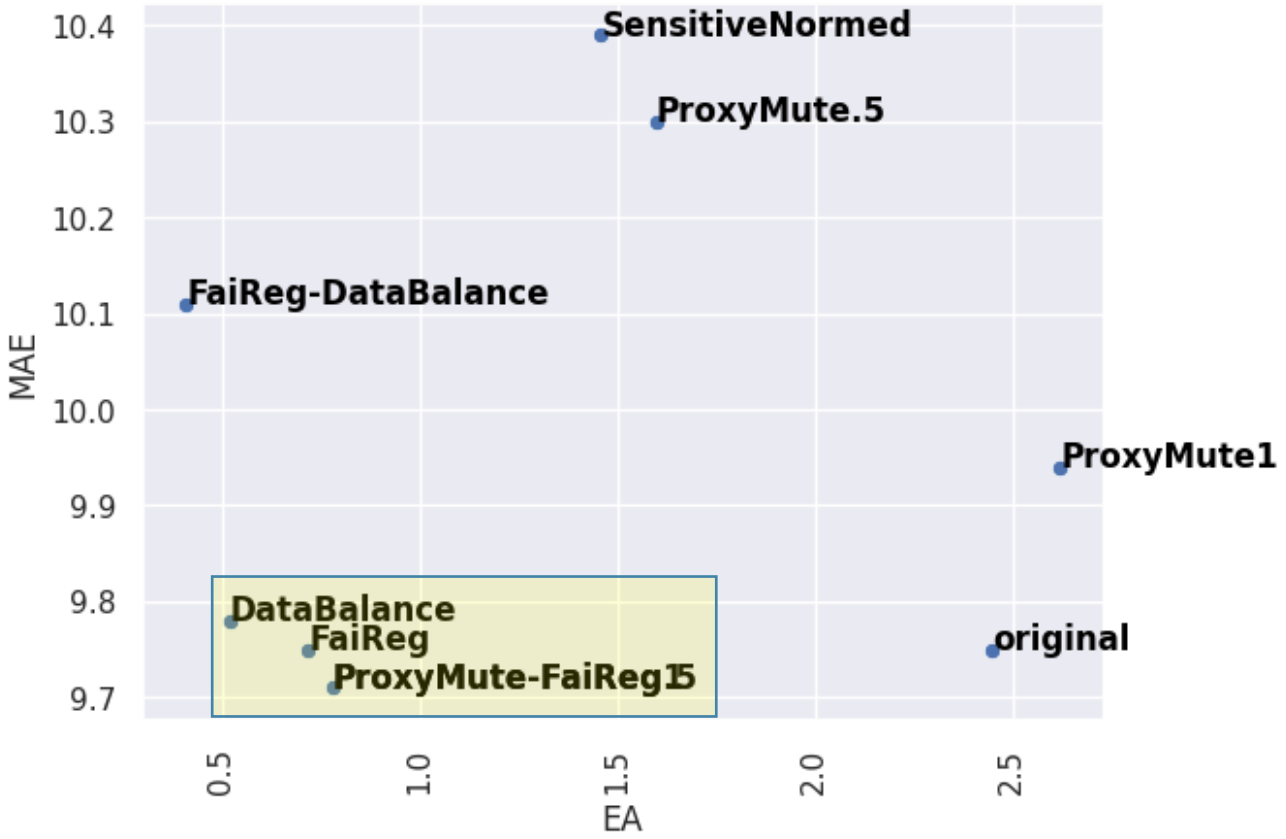
$$Equal\_Accuracy = Mean\_Absolute\_Error\left(Y, \hat{Y} \middle| A = 0\right) - Mean\_Absolute\_Error\left(Y, \hat{Y} \middle| A = 1\right),$$

where A denotes the sensitive attribute.

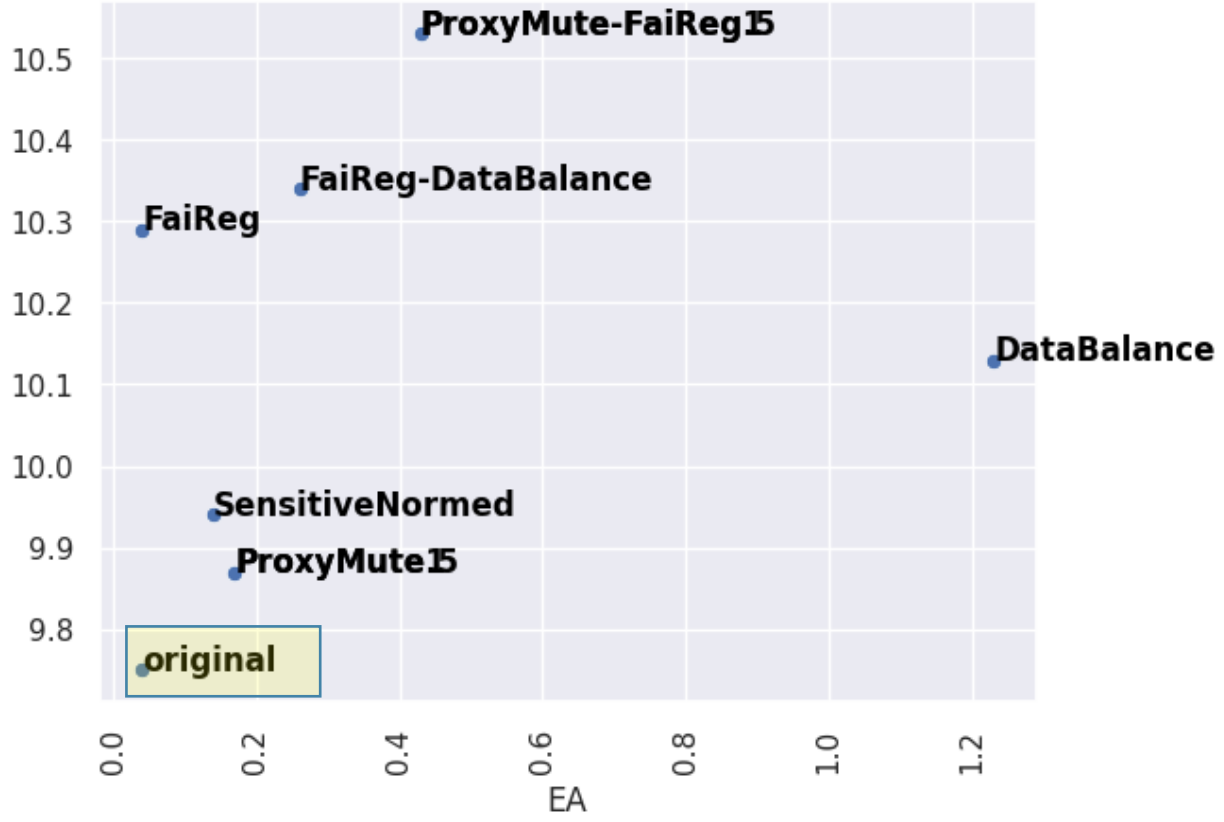
[A] Kaya et al. "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs.", 2017, CVPRW

ChaLearn: Comparison of methods for the multi-modal model

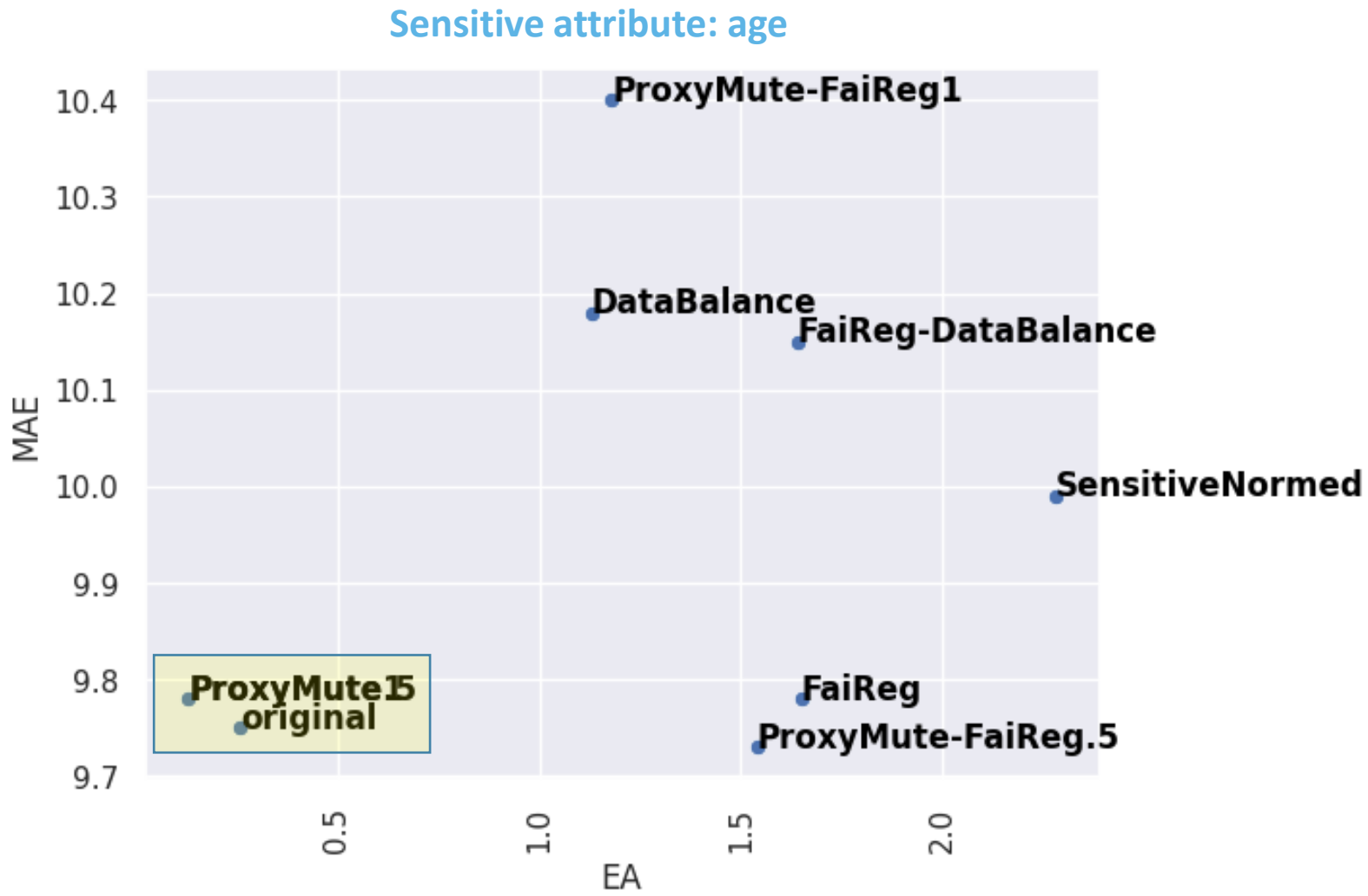
Sensitive attribute: race



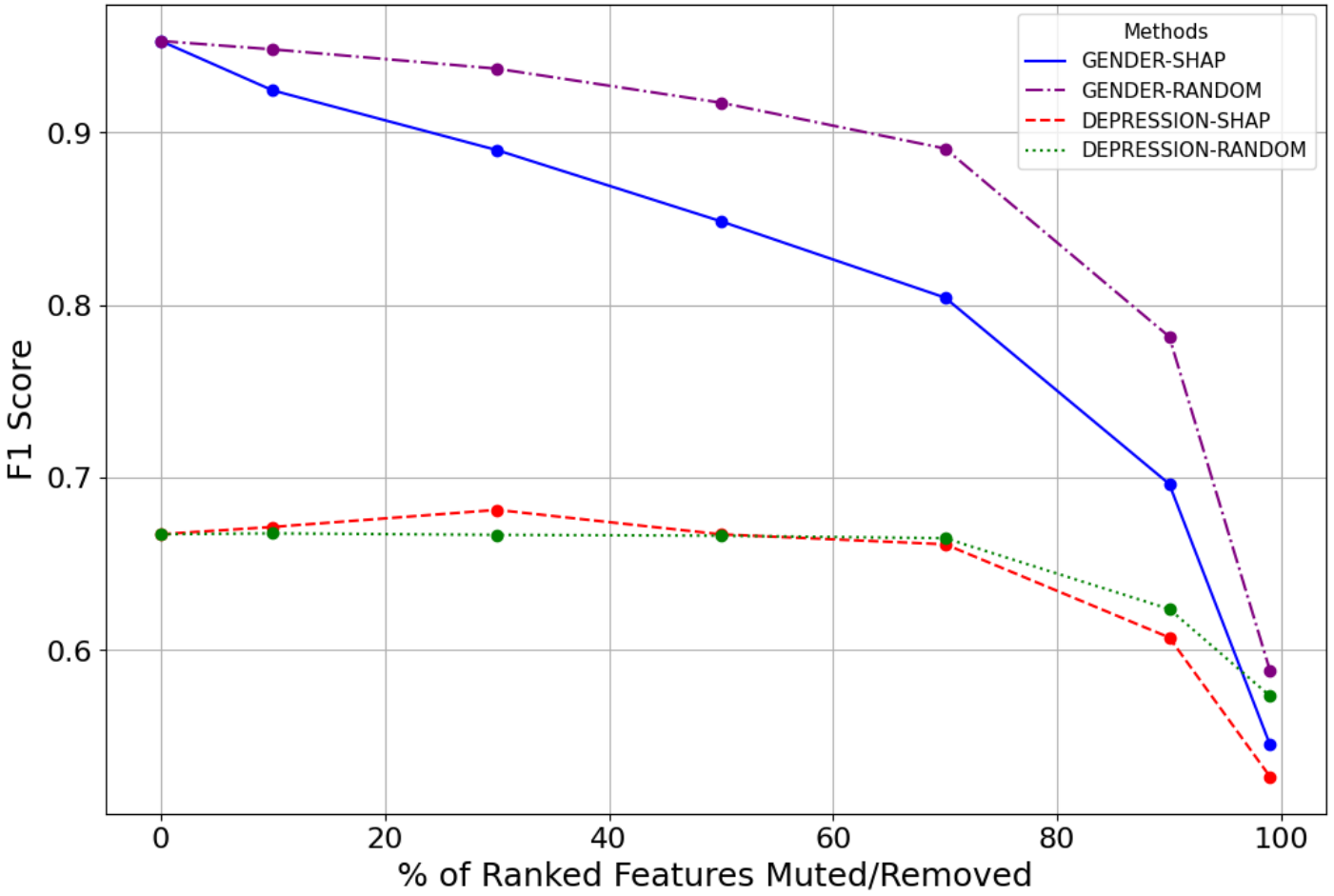
Sensitive attribute: sex



ChaLearn: comparison of methods for the multi-modal model



Application to Depression Phenotype Recognition



On MIMIC-III clinical notes dataset

- Reduced bias without impacting accuracy

Summary

ProxyMute

- **Contribution:** We introduce a novel bias mitigation method using a feature attribution-based explainability approach.
- **Advantage**: no re-training, can be applied for any supervised task with suitable XAI methods.
- **Limitations**: The approach is based on
  1. two important assumptions: **feature independence, feature linearity,**
  2. quantitative fairness measures.

What are the fairness perceptions of clinicians for specific use-cases in the mental health domain?

Use-cases

- Depression phenotype recognition
- Inpatient violence prediction

Sogancioglu, G. et al. (2024). Fairness in AI-based mental health: Clinician perspectives and bias mitigation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.*

Results of semi-structured interviews

- Conducted between March 20, 2024, and April 11, 2024.
- Duration of interviews: 20-30 minutes.
- There were 8 participants in total (6 psychiatrists, 2 clinical psychologists).
- Five interviews were conducted for the depression use-case, and six for the violence use-case.

Sogancioglu, G. et al. (2024). Fairness in AI-based mental health: Clinician perspectives and bias mitigation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.*

Identified Themes Through Manual Analysis

1) Use-Case/Goal/Intervention dependent fairness
2) No sacrifice from accuracy
3) No fairness through unawareness
4) Variable importance of performance measures by gender
5) Awareness of gender biases in clinical practices
6) Communicating model limitations/bias to clinicians
7) Fairness beyond gender

Sogancioglu, G. et al. (2024). Fairness in AI-based mental health: Clinician perspectives and bias mitigation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.*

To Conclude

- Interpretability
  - becomes a *must-have* in mental health domain,
  - can be provided over clinically motivated concepts,
  - is necessary but not sufficient for responsible AI & deployment.
- Algorithmic bias
  - measure needs to be tailored for each use-case and even sensitive group,
  - may remain hidden due to unavailability of sensitive attributes,
  - can be detected and mitigated via XAI tools.
- Privacy preservation is also extremely important
  - Federated Learning [A]
  - Differential Privacy [B]

[A] Borger, T. et al. (2022). Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. *Expert Systems with Applications.*
[B] van der Steen, F., Vink, F., & Kaya, H. (2025). Privacy constrained fairness estimation for decision trees. *Applied Intelligence*

Last but not the least: questions for you!

1. How can we collect/model patient text/speech data to predict **future treatment response** accurately?
2. How that can be done in a responsible, namely, **privacy preserving, fair and interpretable,** manner?
3. Can we use Large Language Models for **interpretable and privacy preserving** manner for this task? Why/not?

Questions

- Thank you for your patience!

Contact: Heysem Kaya <h.kaya@uu.nl>