Exploring Fusion Techniques in Multimodal AI-Based Recruitment: Insights from FairCVdb

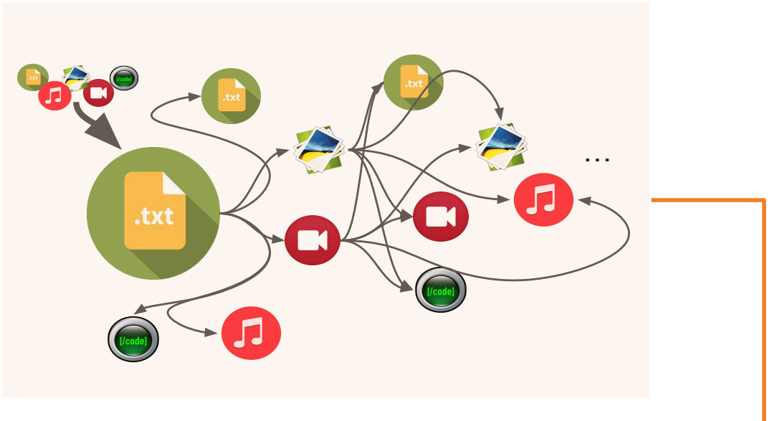
Paper ID: 68

Swati Swati[1] , **Arjun** Roy[1,2] and **Eirini** Ntoutsi[1]

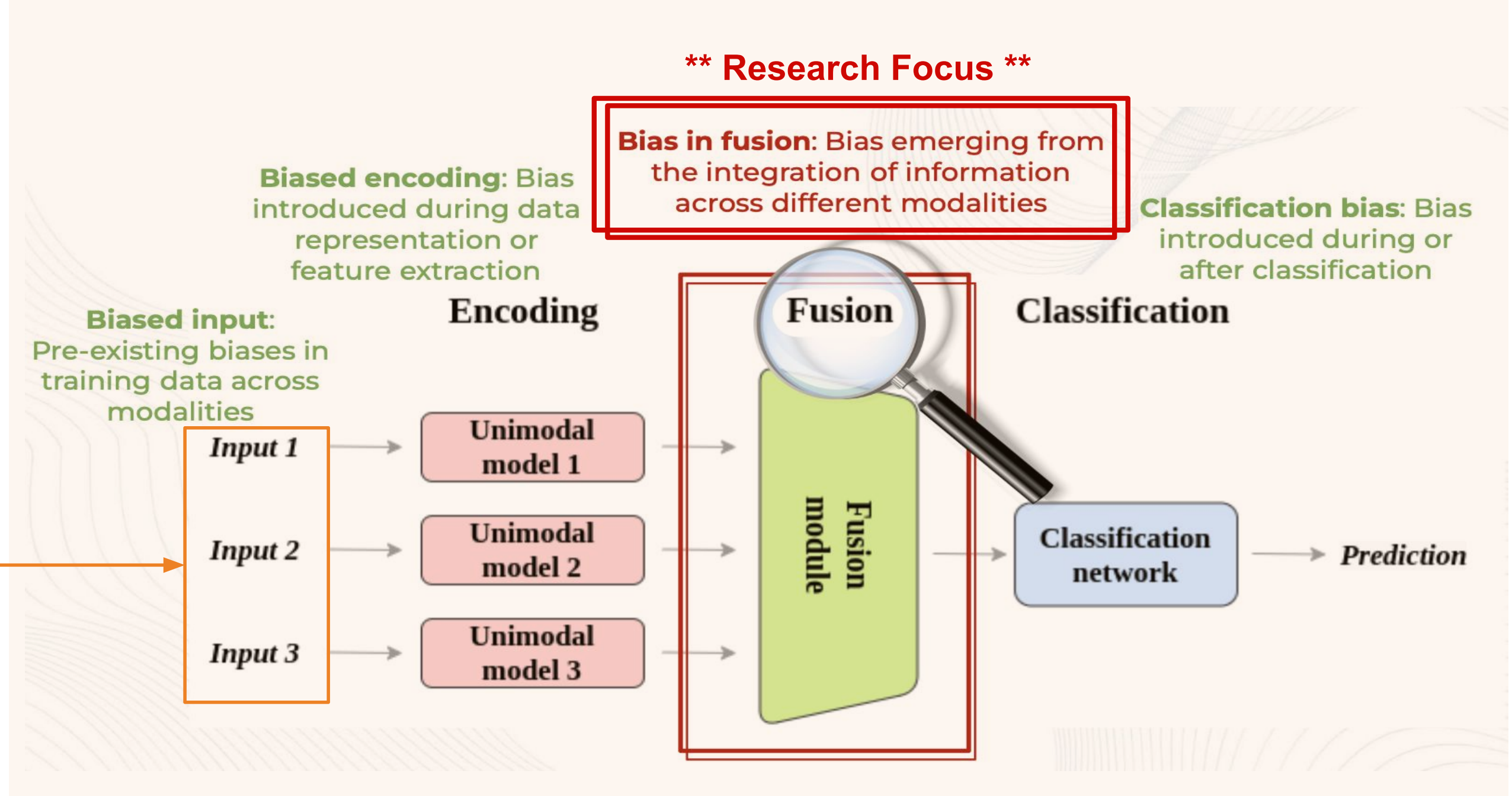
[1]Research Institute CODE, University of the Bundeswehr Munich, Germany, [2]Institute of Computer Science, Free University Berlin, Germany

● INTRODUCTION & MOTIVATION ●

Research Objective: Investigate the **fairness and bias implications of Fusion Approaches** in multimodal AI systems.



Multimodal learning integrates data from different modalities such as images, text, tabular data to enhance decision making.


** Research Focus **
Bias in fusion: Bias emerging from the integration of information across different modalities
Biased encoding: Bias introduced during data representation or feature extraction
Classification bias: Bias introduced during or after classification
Biased input: Pre-existing biases in training data across modalities
Encoding
Fusion
Classification
Input 1
Input 2
Input 3
Unimodal model 1
Unimodal model 2
Unimodal model 3
Fusion module
Classification network
Prediction

Bias across stages of multimodal learning.

Real-World Application: Multimodal AI-based **recruitment systems**:

700+ companies are using AI-based recruitment systems[1].

- With the increasing use of multimodal decision-making algorithms, there are rising concerns about transparency and discrimination, especially affecting specific social groups.
- Majority of existing fairness-aware learning approaches focus on single modalities (tabular, images, or text), but there is a need to understand bias and discrimination in a multimodal context.

[1] Harwell, Drew. A face-scanning algorithm increasingly decides whether you deserve the job. *Ethics of Data and Analytics*. 2022.

● EXPERIMENTAL SETUP ●

Dataset: FairCVdb[2] for fairness study:

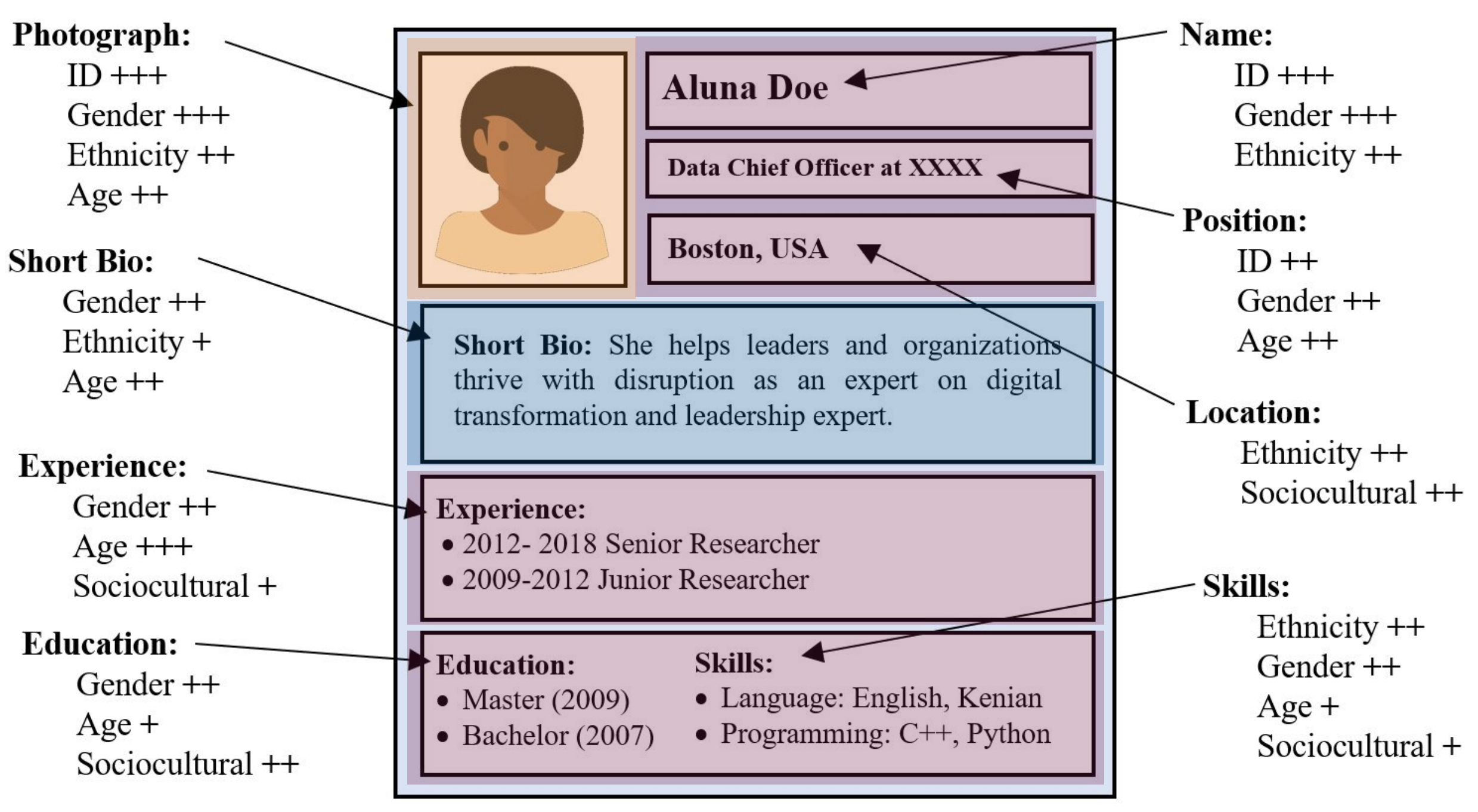
- **Synthetic** research dataset: **24,000** profiles which contain **rich multimodal information** tailored to assess fairness and bias aspects in AI-driven recruitment algorithms.
- **Modalities**: **Visual** (image), **Tabular** (attributes from US Census 2018 Education Attainment data), **Textual** (short bio).
- **Protected attributes**: **Gender**: Female, Male. **Ethnicity**: Asian, Caucasian, African-American.

Task: Determine the subject's probability to be called for a job interview.

Methodology: Recruitment model to predict scores based on candidate resumes, following the methodology from Peña et al. (2023)[3].

Evaluation Metrics:

- Mean Absolute Error (MAE) to measure prediction error.
- Kullback-Leibler (KL) divergence to measure differences between demographic distributions.



Photograph:
ID +++
Gender +++
Ethnicity ++
Age ++

Short Bio:
Gender ++
Ethnicity +
Age ++

Experience:
Gender ++
Age +++
Sociocultural +

Education:
Gender ++
Age +
Sociocultural ++

Aluna Doe

Data Chief Officer at XXXX

Boston, USA

Short Bio: She helps leaders and organizations thrive with disruption as an expert on digital transformation and leadership expert.

Experience:
- 2012- 2018 Senior Researcher
- 2009-2012 Junior Researcher

Education:
- Master (2009)
- Bachelor (2007)

Skills:
- Language: English, Kenian
- Programming: C++, Python

Name:
ID +++
Gender +++
Ethnicity ++

Position:
ID ++
Gender ++
Age ++

Location:
Ethnicity ++
Sociocultural ++

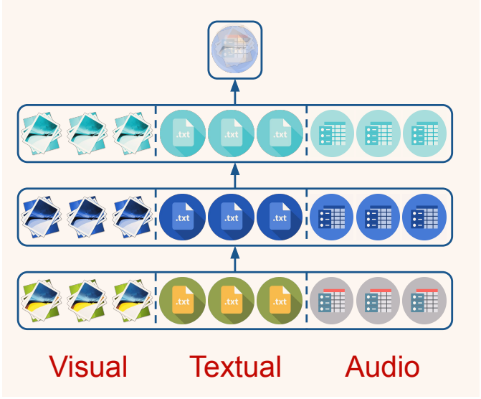
Skills:
Ethnicity ++
Gender ++
Age +
Sociocultural +

Information blocks in a resume and personal attributes that can be derived from each block. The number of crosses represent the level of sensitive information (+++ = high, ++ = medium, + = low)

Multimodal Fusion Strategies: Early and Late:

Early Fusion (Feature-Level Fusion):

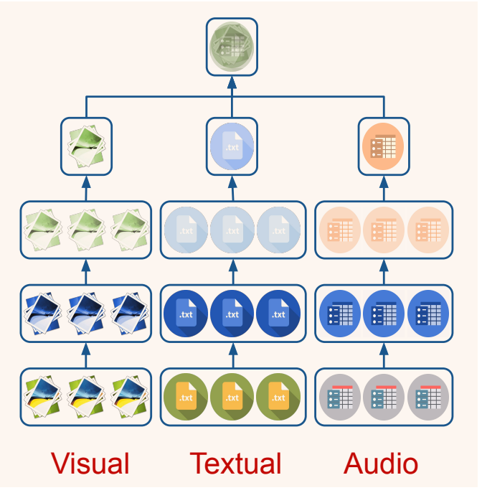
- Occurs at beginning, typically before the data is fed into a neural network.
- Advantageous when inter-modality relationships are simple.



Early fusion

Late Fusion (Classifier-Level Fusion):

- Occurs at the final decision-making stage, after each modality has been processed separately and the decision scores have been calculated.
- Advantageous when modalities exhibit highly distinct data characteristics.



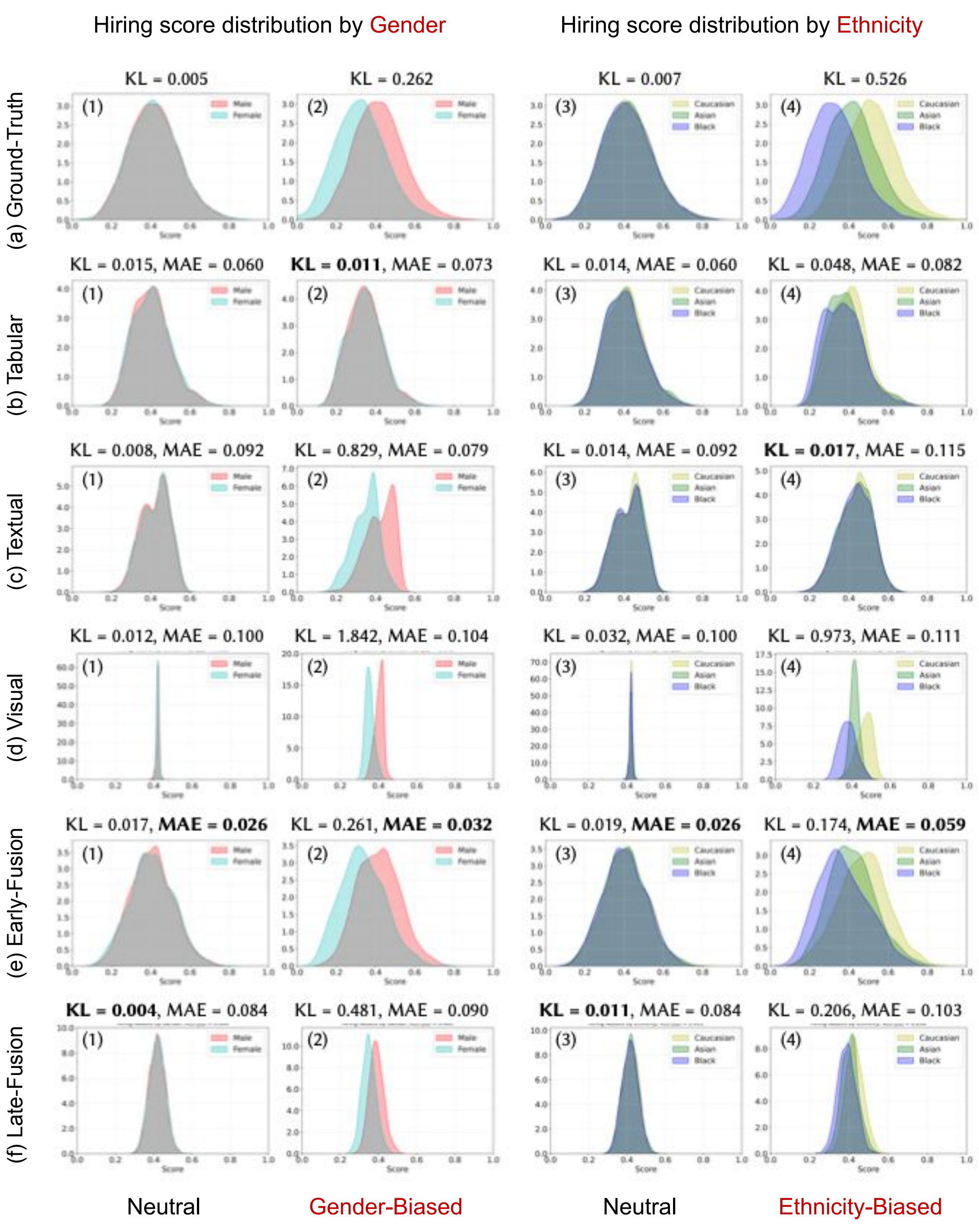
Late fusion

[2] Pena, Alejandro, et al. "Bias in multimodal AI: Testbed for fair automatic recruitment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
[3] A. Peña, I. Serna, A. Morales, J. Fierrez, A. Ortega, A. Herrarte, M. Alcantara, J. Ortega-Garcia, Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment, SN Computer Science 4 (2023) 434.

● RESULTS & CONCLUSIONS ●

KL-divergence between hiring score distributions across Gender and Ethnicity demographics:

- **Neutral**: Unbiased Ideal-World Scenario:
  - Ground-truth: closely aligned for both demographics.
  - Tabular: exhibits a lower score distribution centered around a mean of 0.4 with a negatively-skewed distribution, underestimating the ground-truth.
  - Textual: shows a bimodal distribution, differentiating between instances with high and low scores.
  - Visual: concentrates the distribution within a narrow range [0.39–0.44], indicating over-generalization of the mean score.
  - Late-fusion: produces the least biased results for both demographics, but affected by the extremity of the visual modality, leading to over-generalization and higher MAEs.
  - Early-fusion: delivers the most accurate predictions with the lowest MAEs by effectively learning and resolving unique peculiarities of each modality, such as underestimation, over-generalization, and bimodal distribution, resulting in a shape that closely resembles the ground-truth.
- **Biased**: Biased Real-World Scenario:
  - Ground-truth: unaligned for both demographics.
  - Tabular: exhibits underestimation across all demographics, leading to close alignment of demographic-specific distributions.
  - Textual: shows a favorable skewness for males in job-related words, but no such bias is observed regarding ethnicity, suggesting higher gender-skewness than ethnicity-skewness.
  - Visual: demonstrates the most extreme bias for both demographics, showing bias towards males for gender and overgeneralizing Asians, discriminating against Blacks, and favoring Caucasians.
  - Early-fusion: mimics the ground-truth for both demographics, yielding the lowest MAEs while maintaining fairness.
  - Late-fusion: tends to over-generalize the mean score, resulting in higher MAEs and KL scores.



KL-divergence, MAE, and score distributions across Gender and Ethnicity demographics for different modalities and bias setups. Interpretation: A smaller KL-divergence indicates better alignment between distributions, implying less bias, while a lower MAE indicates a smaller margin of error.

Key Conclusions:

- Fusion techniques play a crucial role in addressing fairness and bias in multimodal AI. Nonetheless, they have the potential to amplify biases from individual modalities, and blindly fusing them may not lead to optimal results.
- *Early fusion* closely mimics ground truth for both demographics and achieves lowest MAEs by incorporating unique characteristics of each modality effectively. It yields fairer solutions even in the presence of demographic biases.
- *Late fusion* leads to highly over-generalized mean scores, resulting in higher MAEs.

Future Directions:

- Bias-aware fusion strategies: Mid-fusion may enhance fairness and accuracy by strategically selecting and combining modalities.
- Test the applicability of these findings across diverse datasets and domains beyond hiring for broader impact and relevance.

Ethics statement: Understanding the risks of using simulated or synthetic data is crucial for fairness, transparency, and effectiveness in automated hiring processes.

For code and additional insights, visit: https://github.com/Swati17293/Multimodal-AI-Based-Recruitment-FairCVdb **or write to:** swati.swati@unibw.de

Funded by the European Union

Co-funded by the European Union

This research work is funded by the European Union under the Horizon Europe MAMMOth project, Grant Agreement ID: 101070285. UK participant in Horizon Europe Project MAMMOth is supported by UKRI grant number 10041914 (Trilateral Research LTD). The research is also supported by the EU Horizon Europe project STELAR, Grant Agreement ID: 101070122.