IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

MAG-BERT-ARL for Fair Automated Video Interview Assessment

BIMASENA PUTRA[1], KURNIAWATI AZIZAH[1], (Member, IEEE), CANDY OLIVIA MAWALIM[2], (Member, IEEE), IKHLASUL AKMAL HANIF[1], SAKRIANI SAKTI[3], (Member, IEEE), CHEE WEE LEONG[4], SHOGO OKADA[2], (Member, IEEE)

[1]Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia
[2]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[3]Graduate School of Science and Technology, Nara Institute Science and Technology, Ikoma 630-0192, Japan
[4]Educational Testing Service, Princeton, NJ 08541, USA

Corresponding authors: Bimasena Putra (bimasena.putra@ui.ac.id), Kurniawati Azizah (kurniawati.azizah@cs.ui.ac.id), and Candy Olivia Mawalim (candylim@jaist.ac.jp)

ABSTRACT Potential biases within automated video interview assessment algorithms may disadvantage specific demographics due to the collection of sensitive attributes, which are regulated by the General Data Protection Regulation (GDPR). To mitigate these fairness concerns, this research introduces MAG-BERT-ARL, an automated video interview assessment system that eliminates reliance on sensitive attributes. MAG-BERT-ARL integrates Multimodal Adaptation Gate and Bidirectional Encoder Representations from Transformers (MAG-BERT) model with the Adversarially Reweighted Learning (ARL). This integration aims to improve the performance of underrepresented groups by promoting Rawlsian Max-Min Fairness. Through experiments on the Educational Testing Service (ETS) and First Impressions (FI) datasets, the proposed method demonstrates its effectiveness in optimizing model performance (increasing Pearson correlation coefficient up to 0.17 in the FI dataset and precision up to 0.39 in the ETS dataset) and fairness (reducing equal accuracy up to 0.11 in the ETS dataset). The findings underscore the significance of integrating fairness-enhancing techniques like ARL and highlight the impact of incorporating nonverbal cues on hiring decisions.

INDEX TERMS Automatic video interview assessment, fairness, model interpretability, adversarial learning

I. INTRODUCTION

In recent years, there has been an increasing trend in the use of automated video interview assessment systems for evaluating candidates across various industries [1]. These systems offer several benefits, including saving time for hiring managers and providing flexibility for candidates in scheduling interviews [2]. Today, automated video interview assessment systems are used worldwide. Notably, HireVue, a prominent company in the automated hiring industry, claims to have assessed over 80 million interviews [3].

Automated video interview assessment systems offer a convenient and efficient method for candidate evaluation in today's globalized and remote work environments. However, the increasing reliance on these systems has raised concerns regarding fairness and the collection of sensitive data, potentially disadvantaging individuals from underrepresented groups [4]. Ethical and legal issues, particularly anti-discrimination laws, must be navigated to prevent the embedding of biases that could perpetuate discrimination, such as gender bias [1], [5], [6]. The General Data Protection Regulation (GDPR) strictly regulates the protection of sensitive data such as age, gender, and race, in accordance with the EU Data Protection Directive [7]–[9]. As a result, some existing research [10]–[12] falls short in terms of addressing fairness without using sensitive data. Additionally, under the EU's AI Act [13], automated interview systems are classified as high-risk AI, necessitating stringent observability, human oversight, and transparency requirements [14]. Consequently, the development of fair and interpretable automated video interview assessment systems that do not rely on sensitive attributes has become a critical area of research and development.

Addressing these issues requires a solution that ensures fairness without relying on sensitive attributes. Previous research in automated video interviews, including research by [11], [15], and [16], has attempted to mitigate bias. However, these approaches still depend on sensitive attributes to achieve fairness. In contrast, [17] developed automated video interview systems that do not incorporate sensitive attributes, but their methods fail to adequately address fairness concerns.

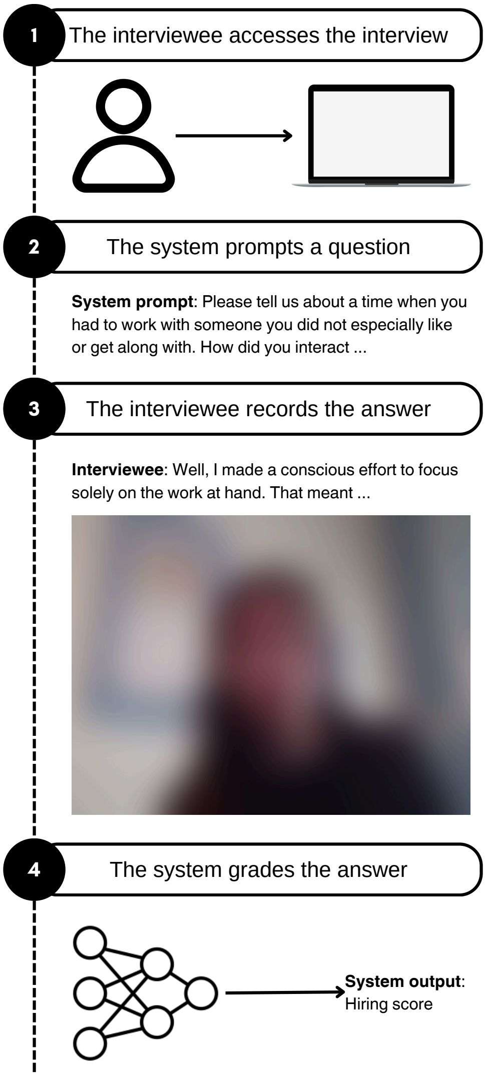
This research introduces MAG-BERT-ARL to ensure fairness in automated video interview assessment systems without relying on sensitive attributes. MAG-BERT-ARL combines Multimodal Adaptation Gate and Bidirectional Encoder Representations from Transformers (MAG-BERT) [18], which has demonstrated effectiveness in incorporating nonverbal cues with text, with Adversarially Reweighted Learning (ARL) [19], a technique designed to enhance Rawlsian Max-Min Fairness without demographics. In this context, Rawlsian fairness refers to optimizing accuracy for the least advantaged groups [19]. Several modifications were incorporated to enhance its effectiveness. While ARL is traditionally applied to tabular data, this adaptation leverages the CLS token from BERT, thus integrating multiple modalities from input videos and addressing fairness concerns. The approach employs techniques such as batch normalization [20], skip connections [21], and ReLU activation functions [22] to enhance performance and fairness. These methods are important in preventing the vanishing gradient problem that could arise from integrating MAG-BERT with ARL. Furthermore, Gradient SHAP [23] is utilized to assess the contribution of individual input features, providing an interpretable evaluation of the model's predictions.

Experiments were conducted on the Educational Testing Service (ETS) [17] and First Impressions (FI) [24] datasets. Three variants of MAG-BERT-ARL were developed, each utilizing different loss functions: Mean Squared Error (MSE), Binary Cross-Entropy (BCE), and their combination. ARL integration improved both fairness and performance, with the MSE variant excelling on the ETS dataset and the combined loss function variant performing best on the FI dataset. Additionally, ablation tests and model interpretations were performed on text, acoustic, and visual modalities, indicating that a performance-fairness trade-off exists when incorporating nonverbal modalities. For instance, the trade-off is evident when comparing the use of only the text modality against the combined use of text, acoustic, and visual modalities.

In summary, the contributions are as follows:

1) Propose MAG-BERT-ARL to address fairness in automated video interview assessment systems without relying on sensitive data,
2) Improve the ARL approach as an automatic debiasing method in the multimodal interview assessment domain.

The rest of this paper is organized as follows: Section II explains previous related works, Section III introduces MAG-BERT-ARL, the proposed method for assessing video interviews. Section IV presents the experiment details of this research. Section V presents the results and analyses. Section VI presents the summary.


1
The interviewee accesses the interview
2
The system prompts a question
3
The interviewee records the answer
4
The system grades the answer
System output: Hiring score

System prompt: Please tell us about a time when you had to work with someone you did not especially like or get along with. How did you interact ...

Interviewee: Well, I made a conscious effort to focus solely on the work at hand. That meant ...

FIGURE 1. **Automated Video Interview Assessment Systems in Practice**

II. RELATED WORKS

This section presents advancements in automating video interviews, fair automated video interview assessment systems, and fairness without demographics methods.

A. AUTOMATED VIDEO INTERVIEW ASSESSMENT SYSTEMS

The development of automated video interview assessment systems began when [25] developed a system for automatic personality recognition using support vector machines (SVM) and lexical features. Nguyen et al. [26] developed a system for inferring hirability in face-to-face interviews for marketing jobs that utilizes nonverbal cues. However, it was not until 2016 that these systems were adapted for asynchronous video interviews [27].

In asynchronous video interview systems, hiring managers can configure questions for the system to display on the candidate's screen [28]. During the interview, candidates are prompted to answer these questions, with some preparation time allotted [29]. The candidates' responses are recorded via camera. Subsequently, the system assesses their answers and personalities, assigning a score that determines whether the candidate will advance to the next hiring stage. This mechanism is illustrated in Fig. 1.

Several previous studies on automated video interview assessment systems have utilized different architectures and methods for embedding and fairness mitigation. Singhania et al. [15] introduced a technique for assessing video interviews with fairness considerations, employing linear regression with L-2 regularization. Similarly, [11] proposed a method utilizing data balancing and adversarial learning to mitigate bias in the same context. Furthermore, [16] presented an approach leveraging adversarial learning and Wasserstein distance regularization to address fairness concerns in automated video interview assessment systems. Specifically, models developed by [11], [17], [30], and [16] have shown significant promise in this domain. These studies have been tested on ETS and FI datasets. Unlike these previous studies, our research aims to develop an automated video interview assessment system that does not discriminate and does not require sensitive attributes. While [11] and [16] focus on mitigating bias in automated video interview assessment, our research focuses on enhancing fairness without utilizing sensitive attributes.

Chen et al. [17] utilized text and acoustic features for their model, employing bag-of-words feature extraction for text and PyAudioAnalysis for acoustic features. In contrast, research by [11] and [15] focused only on acoustic and visual features. Yan et al. [11] used openSMILE for acoustic features and VGG-Face, LGBP-TOP, and VGG-VG-19 for visual features. Singhania et al. [15] applied openSMILE for acoustic features and OpenFace's Head Pose and Facial Action Unit for visual features. Additionally, [30] and [16] incorporated all three modalities: text, acoustic, and visual features. Rahman et al. [30] utilized the ALBERT model for text and a multiple instance learning model for both acoustic and visual features. Kim et al. [16] combined BERT for text, PyAudioAnalysis for acoustic, and CCN-LSTM for visual features. In our research, text features are extracted using BERT [31], acoustic features are obtained through eGeMAPS [32], and visual features are derived from OpenFace [33] to maximize both performance and fairness.

Chen et al. [17] employed a unified method utilizing clustering and text classification techniques. This model has demonstrated effectiveness in predicting hiring recommendation scores without utilizing sensitive attributes. However, it does not address fairness issues in automated video interview assessment systems. Rahman et al. [30] introduced an approach leveraging integrated gradients [34] and Gradient SHAP [23], which significantly improved the interpretability of the system. Despite these advancements, their model also does not address fairness issues in automated video interview assessment systems. Research by [11], [15], and [16] proposed debiasing approaches, achieving substantial improvements in fairness metrics. However, these models utilize sensitive attributes for their approaches.

As reported by [11], [15] and [16], the performance of their models is sufficient but still not entirely satisfactory, particularly due to their reliance on sensitive attributes for fairness mitigation. The baseline model in our research adopts MAG-BERT [18]. To improve fairness performance, our research incorporates Adversarially Reweighted Learning (ARL) [19] to the baseline MAG-BERT model with some modifications in three categories: input type, training strategy, and activation function type.

B. FAIRNESS WITHOUT DEMOGRAPHICS

To address the ethical and legal obstacles, numerous research has attempted to mitigate bias by eliminating sensitive attributes like gender and race while upholding fairness within the system. One approach to achieving this is through methods such as fairness without demographics. Examples of such methods include Distributionally Robust Optimization (DRO) [35], which employs distributionally robust optimization to handle worst-case scenarios across groups, Adversarially Reweighted Learning (ARL) [19], which focuses on identifying training errors through an adversarially reweighted learning approach, Shared Latent Space-Based Debiasing (SLSD) [36], which uses adversarial learning to debias latent representations, and transformer without demographics [37], which focus on debiasing transformers.

Our research employs Adversarially Reweighted Learning (ARL) [19] to promote fairness without using sensitive attributes. Unlike ARL and similar methods such as DRO [35] and SLSD [36], which are typically applied to tabular data like financial information, our research proposes a method to apply fairness without demographics techniques in the video domain. By integrating ARL into MAG-BERT, the research develops MAG-BERT-ARL. This approach extends beyond transformers without demographics [37], which have been applied in text and visual modalities, by also incorporating acoustic modality.

III. METHODS

This research presents an approach that can be applied within regulatory boundaries. Given the constraint of not having access to sensitive attribute, addressing fairness becomes challenging. In this context, Adversarially Reweighted Learning (ARL) [19] is utilized as the standard Multi-Layer Perceptron (MLP) head due to its ability to achieve fairness without demographics. For the baseline model, MAG-BERT, a model rooted in the Multimodal Augmentation Gate (MAG) [18], is utilized. Some changes are applied to ensure the integration of these two components.

First, traditionally, ARL takes tabular data as input; here, it is repurposed to utilize the CLS token, known to be the primary input for the classifier in a typical BERT model [38].

TABLE 1. Summary of studies with their respective datasets, tasks, debiasing methods, information about the fairness and the usage of sensitive data, and metrics. Fair-aware indicates whether the model incorporates a debiasing method, marked as true (v) if it does and false (x) if it does not. Meanwhile, no sensitive data specifies whether the model utilizes sensitive data, such as age, gender, and race, during training and validation, marked as true (v) if it does not and false (x) if it does.

Authors	Dataset	Task	Debiasing Method	Fair-aware	No Sensitive Data	Metrics
Chen et al. (2017) [17]	ETS	Classification	None	x	v	Precision, Recall, F1
Rahman et al. (2021) [30]	ETS	Classification	None	x	v	Accuracy, AUC Score
Singhania et al. (2020) [15]	Video dataset	Regression	Lasso & Ridge regularization	v	x	Pearson Correlation Coefficient, Mean Difference, Mean Absolute Error, Demographic Parity, Equalized Odds, Equal Opportunity, Disparate Impact
Yan et al. (2020) [11]	FI	Classification	Adversarial learning	v	x	Accuracy, Demographic Parity, Equal Accuracy
Kim et al. (2023) [16]	Hiring recommendation, FI	Regression	Wasserstein regularization, adversarial learning	v	x	Pearson Correlation Coefficient, Spearman's Rank Correlation Coefficient, Strong Pairwise Demographic Disparity, Strong Pairwise Equal Opportunity

This adaptation enables the leverage of rich contextual information encoded by BERT while applying ARL to address fairness concerns.

Second, the vanishing gradient problem can arise when training deep neural networks, particularly in scenarios where the gradient signal diminishes as it propagates backward through layers [39]. To mitigate this issue in MAG-BERT-ARL, techniques such as batch normalization and skip connections are utilized [20], [21]. These methods help alleviate the vanishing gradient problem by facilitating smoother gradient flow during training [40].

Third, the choice of activation function plays a crucial role in the performance of neural networks. In MAG-BERT-ARL, activation functions that facilitate better gradient flow are opted for [22]. In this selection, rectified linear units (ReLU) are preferred over the sigmoid function due to their demonstrated superior performance in specific cases, while also addressing the issue of vanishing gradients.

Finally, to comprehend the individual contributions of input features to a model's predictions and understand the model's functioning, a mathematical framework called Gradient SHAP is utilized [23]. The influence of each input feature on the model's output is assessed by this interpretative approach, emphasizing the isolation of each feature's impact [41]. By using Gradient SHAP, the contributions of each feature to the model's predictions can be evaluated.

A. MODEL ARCHITECTURE

Fig. 2 presents the proposed MAG-BERT-ARL architecture. Given a video interview, MAG-BERT-ARL predicts the output hiring recommendation score $\hat{y}$. MAG-BERT-ARL consists of MAG-BERT [18] and ARL [19]. The video is processed to extract triplets of features: text, acoustic, and visual. These features are passed as inputs through MAG-BERT. MAG-BERT processes the input and produces a representation $\overline{Z} = [\overline{Z}_{\text{CLS}}, \overline{Z}_1, \overline{Z}_2, ..., \overline{Z}_{\text{N}}]$. Subsequently, the CLS token $\overline{Z}_{\text{CLS}}$ is passed through both the learner and the adversary of ARL if it is in training mode (when y exists). The learner outputs logits that are utilized as the model predictions $\hat{y}$, while the adversary generates adversary weights λ, which are utilized for training the ARL's loss function, as in Eq, 8. There are three variants of MAG-BERT-ARL, each utilizing different loss functions: MAG-BERT-ARL M applies the Mean Squared Error (MSE) loss function, MAG-BERT-ARL B applies the binary cross-entropy loss function, and MAG-BERT-ARL MB combines the MSE and binary cross-entropy loss functions.

1) MAG-BERT

MAG-BERT, an extension of the BERT model [31], incorporates the Multimodal Augmentation Gate (MAG) mechanism. MAG processes three types of inputs: textual, acoustic, and visual. Each non-textual feature (*A*, *V*) is combined with the text embedding input *Z* and subjected to attention gating to generate a relevant information vector, denoted as g, as described in Eq, 1.

$$\begin{aligned} g^a &= R(W_{ga}[Z; A] + b_a) \\ g^v &= R(W_{gv}[Z; V] + b_v) \end{aligned} \tag{1}$$

Subsequently, a verbal shift vector, *H*, is formed by integrating non-textual features scaled by their respective vectors g, following Eq, 2.

$$H = g^a \cdot (W_a A) + g^v \cdot (W_v V) + b_h \tag{2}$$

The original text input is then adjusted by a factor α times vector *H*, as outlined in Eq, 3, where α is determined by a scaling formula defined in Eq, 4 where β is a hyperparameter.

$$\alpha = \min(\frac{||Z||}{||H||}\beta, 1) \tag{3}$$

$$\hat{Z} = Z + \alpha H \tag{4}$$

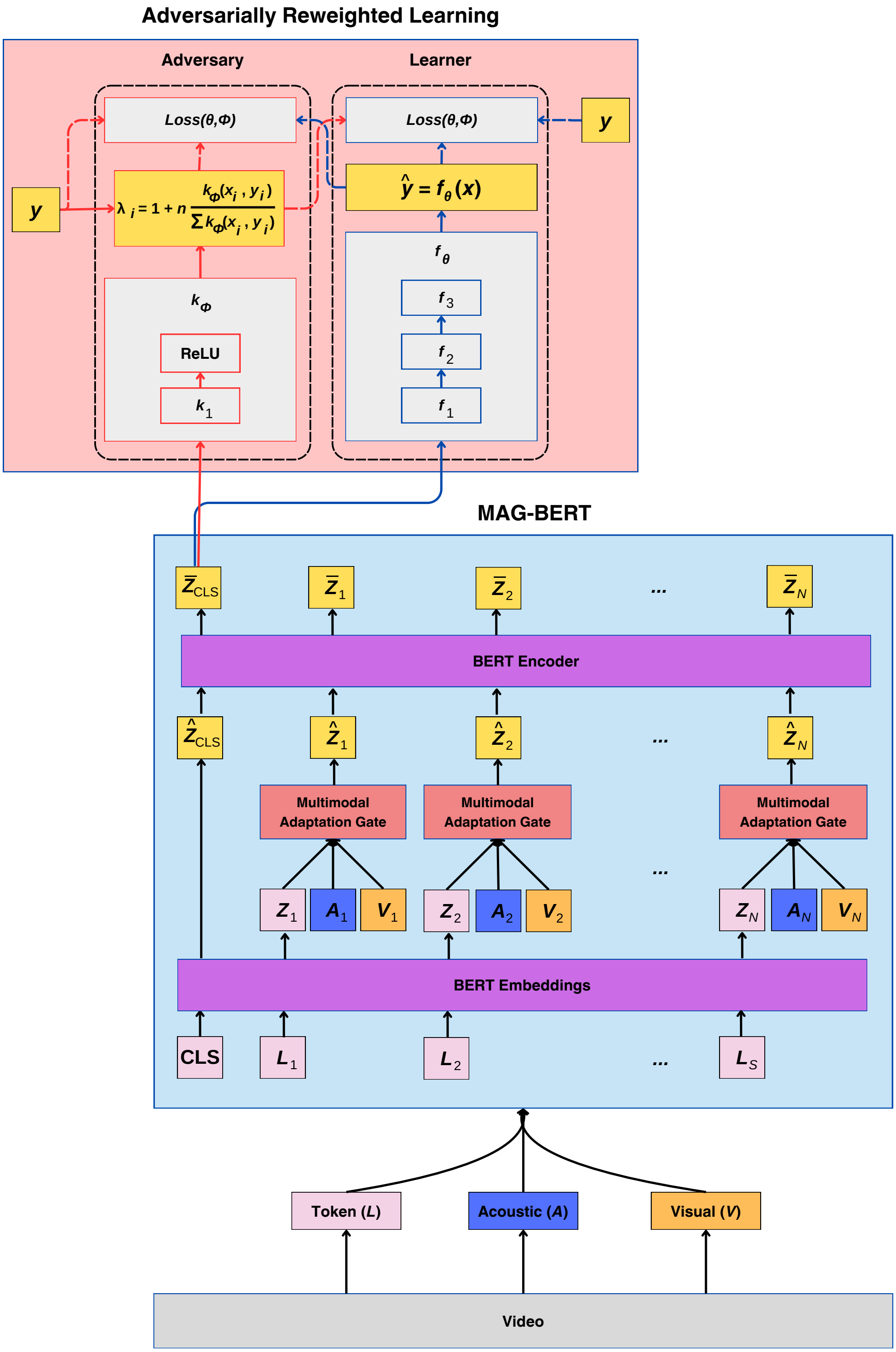


FIGURE 2. **MAG-BERT-ARL architecture. Given a video interview, MAG-BERT-ARL predicts the hiring recommendation score.**

This augmented representation $\hat{Z}$ enriches the initial text input by incorporating not only the spoken words but also associated nonverbal cues.

MAG-BERT extends the MAG framework into BERT. In MAG-BERT, MAG is integrated at a specific layer of BERT, positioned before the BERT encoder and after the BERT embedding. Similar to BERT, the operations within MAG-BERT commence with the token sequence $L = [CLS, L_1, L_2, ..., L_S]$, where CLS is a CLS token utilized for class label prediction. Subsequently, the sequence L

passes through the BERT embedding, producing $Z = [Z_{\text{CLS}}, Z_1, Z_2, ..., Z_{\text{N}}]$, which represents the text embedding. Subsequently, embedding Z is paired with other non-textual feature (A, V), forming a triplet (Z, A, V). This triplet is passed through the MAG which outputs $\hat{Z}$, the unified representation of all multimodalities. Finally, the unified representation $\hat{Z}$ is fed into the BERT encoder, yielding the output $\overline{Z}$.

2) Adversarially Reweighted Learning

Adversarially Reweighted Learning (ARL) consists of two components: an adversary and a learner. The learner, represented by a feed-forward network f_θ, aims to minimize a specified loss function ℓ, typically cross-entropy loss, during training. The learner produces logits $\hat{y}$ by processing the $\overline{Z}_{CLS}$ output from the MAG-BERT, as in Eq, 5.

$$\hat{y} = f_\theta(\overline{Z}_{CLS}) \tag{5}$$

Each feed-forward network within the learner f_θ consists of a linear layer, a dropout layer, a layer normalization, and a ReLU activation function. The network processes an input x through the following steps: initially, x is passed through the linear layer, where it is multiplied by a weight matrix w and summed with a bias vector b, followed by the dropout layer, resulting in $\hat{x}$. This initial input x is also propagated through skip connections, bypassing the linear and dropout layers, and is summed with $\hat{x}$. Both the initial input and the output $\hat{x}$ then undergo post-layer normalization, resulting in $\overline{x}$. Finally, the normalized output $\overline{x}$ is passed through the ReLU activation function, yielding the logits $\hat{y}$, which represent the model's predictions.

Meanwhile, the adversary, also represented by a feed-forward network k_ϕ, seeks to maximize penalties for computationally-identifiable regions with elevated loss by assigning a weight vector $\lambda_\phi(x_i, y_i)$, as shown in Eq, 6, thereby maximizing the loss.

$$\lambda_\phi(x_i, y_i) = 1 + n \cdot \frac{k_\phi(x_i, y_i)}{\sum_{i=1}^{n} k_\phi(x_i, y_i)} \tag{6}$$

Each feed-forward network within the adversary k_ϕ consists of a linear layer, a dropout layer, and a layer normalization. The process mirrors that of the learner. The network processes an input x through the following steps: first, x is passed through the linear layer, where it is multiplied by a weight matrix w and summed with a bias vector b, followed by the dropout layer, resulting in $\hat{x}$. The initial input x is also propagated through skip connections, bypassing the linear and dropout layers, and summed with $\hat{x}$. Both the initial input and the output $\hat{x}$ then undergo post-layer normalization, resulting in $\overline{x}$. This process is repeated at least once. Subsequently, the normalized output $\overline{x}$ is passed through a ReLU activation function, resulting in $\mathcal{X}$. Finally, the output $\mathcal{X}$, along with the target labels y, is used to calculate the adversary weight λ.

B. MODEL TRAINING

The training process for MAG-BERT-ARL involves optimizing the parameters of all its components: MAG-BERT $h_{\theta,\phi}$, the learner f_θ, and the adversary k_ϕ, where θ and ϕ are the parameters that will be optimized. This process comprises two distinct stages. The first stage, referred to as pretraining, involves training only the MAG-BERT and learner components. During this phase, the adversary weight λ does not influence the training process, and the model focuses only on minimizing the prediction loss $J(\theta, \phi)$ as defined in Eq, 7.

$$J(\theta, \phi) = \min_\theta \sum_{i=1}^{n} \ell(h_{\theta,\phi}(x_i), y_i) \tag{7}$$

In the second stage, the adversary is also included in the training process. This stage employs a minimax optimization approach to balance the parameters of the learner and the adversary. The learner aims to minimize the prediction loss, while the adversary seeks to maximize the penalty for regions with high computationally-identifiable loss. In response, the learner adapts its parameters to minimize loss in regions identified by the adversary as having high loss, thereby promoting fairness. This iterative process facilitates the optimization of fairness across different identifiable groups by dynamically adjusting the distribution of training samples and can be formally expressed as shown in Eq, 8. The detailed training procedure for MAG-BERT-ARL is outlined in Algorithm 1.

$$J(\theta, \phi) = \min_\theta \max_\phi \sum_{i=1}^{n} \lambda_\phi(x_i, y_i) \cdot \ell(h_{\theta,\phi}(x_i), y_i) \tag{8}$$

IV. EXPERIMENTS

This section presents details of the data used, data preparation steps, model implementation, performance evaluation metrics, and the approach for classifying absent gender labels and interpreting the model's predictions.

A. DATASET

1) Educational Testing Service Dataset

The Educational Testing Service (ETS) Dataset [17] comprises 1,891 recordings capturing human responses to prompts within a simulated interview environment facilitated by a computer program. Each interviewee provides answers to a predetermined set of questions, with their video recordings subsequently annotated with various labels. Specifically, each interviewee is assigned six labels: a hiring score (hs) and the Big Five personality traits, namely agreeableness (ag), extraversion (ex), conscientiousness (co), emotional stability (em), and openness (op). These labels are assessed by five raters using a 7-point Likert scale, with 1 indicating "strongly disagree" and 7 indicating "strongly agree". For the purposes of this research, only the hiring score (hs) is utilized. The dataset is divided into a training set consisting of 1,519

Algorithm 1 Training Procedure for MAG-BERT-ARL

```
Require: Dataset $((X_t, X_a, X_v) \in X$, Segment ID, $Y)$
Ensure: Trained MAG-BERT $h_{\theta,\phi}$, learner $f_\theta$, and adversary $k_\phi$
Initialize MAG-BERT $h_{\theta,\phi}$ with random parameters
Initialize learner $f_\theta$ with random parameters
Initialize adversary $k_\phi$ with random parameters

Set hyperparameters batch size $m$, epoch $e$, pretrain steps $p$, learning rates of $\eta_{learner}$ and $\eta_{adversary}$

for each epoch in e do
    for each step $s$ and batch $B$ in dataset do
        Extract representations in latent space $h_{\theta,\phi}(B)$ from each batch
        Compute learner output $f_\theta(h_{\theta,\phi}(B)_{\text{CLS}})$ using representation CLS token as input
        if step $s >$ pretrain steps $p$ then
            Compute adversary output $k_\phi(h_{\theta,\phi}(B)_{\text{CLS}})$ using representation CLS token as input
            Compute adversary weight: $\lambda_\phi(x_i, y_i) \leftarrow 1 + n \cdot \frac{k_\phi(x_i,y_i)}{\sum_{i=1}^{n} k_\phi(x_i,y_i)}$
            Compute learner loss: $L_{learner} \leftarrow \sum_{i=1}^{n} \lambda_\phi(x_i, y_i) \cdot \ell(h_{\theta,\phi}(x_i), y_i)$
            Update $\theta$ with gradient descent: $\theta \leftarrow \theta - \eta_{learner} \nabla_\theta \, L_{learner}$
            Compute adversary loss: $L_{adversary} \leftarrow \sum_{i=1}^{n} -\lambda_\phi(x_i, y_i) \cdot \ell(h_{\theta,\phi}(x_i), y_i)$
            Update $\phi$ with gradient descent: $\phi \leftarrow \phi - \eta_{adversary} \nabla_\phi \, L_{adversary}$
        else
            Compute learner loss: $L_{learner} \leftarrow \sum_{i=1}^{n} \ell(h_{\theta,\phi}(x_i), y_i)$
            Update $\theta$ with gradient descent: $\theta \leftarrow \theta - \eta_{learner} \nabla_\theta \, L_{learner}$
        end if
    end for
end for
```

TABLE 2. **Educational Testing Service (ETS) & First Impressions (FI) Datasets Summary**

Description	Dataset	
	ETS	FI
Average Video Duration	2-3 minutes	15 seconds
Total Video Duration	78 hours	±41.6 hours
Annotators	5	2500
Interviewees	260	3000
Gender Annotation	Unavailable	Available
Hiring Score (hs)	$1 \leq \text{hs} \leq 7$, $\text{hs} \in \mathbb{N}$	$0 \leq \text{hs} \leq 1$, $\text{hs} \in \mathbb{R}$
Decision Threshold	5.6	0.5
Train Dataset Size	1215	6000
Test Dataset Size	372	2000
Evaluation Dataset Size	304	2000
Total Dataset Size	1891	10000

recordings and a test set containing 300 recordings. Additionally, a validation dataset is created by extracting the last 20% of the training dataset and the remaining 80% becomes the training dataset. The summary of the ETS dataset is presented in Table 2.

2) First Impressions Dataset

The First Impressions (FI) Dataset [24] comprises 10,000 video segments sourced from over 3,000 distinct YouTube videos featuring individuals speaking in English. These video clips are annotated with various attributes including gender, ethnicity, Big Five personality traits, and a continuous variable known as the interview score, which indicates the likelihood of the subject being invited for a job interview. These annotations are assessed on a scale ranging from 0 to 1. Notably, for the purposes of this research, only the interview score is utilized. The dataset is partitioned into three subsets: a 60% training set, a 20% validation set, and a 20% testing set. The summary of the FI dataset is presented in Table 2.

B. FEATURE EXTRACTION

Many modalities can be extracted from a video. Experiments were conducted on the baseline MAG-BERT model [18] with various combinations of features, including eGeMAPS acoustic feature sets [32], OpenFace visual subfeature sets [33]—such as gaze, facial landmarks, action units, and all feature sets—and PyAudioAnalysis acoustic feature sets [42], to identify the optimal configuration for a robust and nuanced evaluation of candidates during automated interviews. From the experiments, eGeMAPS and OpenFace have shown to yield the best performance in trade-offs between performance and bias and both are utilized to extract acoustic and visual features, respectively. For the text features, whisper-timestamped [43]–[45] is utilized.

1) Text Features

whisper-timestamped [43]–[45] is an extension of the Whisper Automated Speech Recognition (ASR) system, incorporating word-level timestamp predictions for each transcribed word. When used for audio transcription, Whisper-timestamped generates an output containing both text and

TABLE 3. eGeMAPS Feature Index

Index (i)	Feature	Length
0-9	Frequency (F0)	10
10-19	Loudness	10
20-21	Spectral	2
22-29	MFCC	8
30-31	Jitter	2
32-33	Shimmer	2
34-35	HNR	2
36-39	Logarithmic Harmonic	4
40-57	Formant (F1-3)	18
58-59	Alpha Ratio	2
60-61	Hammarberg	2
62-65	Slope Voiced	4
66-67	Spectral Flux	2
68-75	MFCC Voiced	8
76	Alpha Ratio Unvoiced	1
77	Hammarberg Unvoiced	1
78-79	Slope Unvoiced	2
80	Spectral Flux Unvoiced	1
81-86	Voiced Segment	6
87	Equivalent Sound Level	1
Total		**88**

corresponding speech segments. Each segment includes a list of words with associated timestamps, accurate to two decimal places, along with a confidence score for each word.

2) Acoustic Features

eGeMAPS [32] is a minimalistic acoustic feature set that serves as a basic standard acoustic parameter set for various areas of automatic voice analysis. eGeMAPS contains 88 parameters of acoustic features, such as frequency, spectral, loudness and pitch. A detailed breakdown of these 88 acoustic parameters is presented in Table 3.

3) Visual Features

OpenFace [33] is a free-to-use toolkit to extract facial features for facial behaviour analysis. OpenFace is able to extract facial features for facial landmark location, head pose, face shape, eye gaze, and facial action unit. From OpenFace, 340 parameters of facial landmark location, 6 parameters of head pose, 40 parameters of face shape, 288 parameters of eye gaze, and 35 parameters of facial action unit can be extracted, totaling 709 parameters. A detailed breakdown of these parameters is presented in Table 4.

C. PREPROCESSING

Prior to inputting the dataset into the machine learning model, it is necessary to preprocess the data into a compatible format. This involves utilizing whisper-timestamped [43]–[45] to transcribe the interviewee's speech from audio recordings into text, while also extracting timestamps corresponding to each word uttered in the speech. These timestamps serve as markers, facilitating subsequent alignment of spoken words with their associated visual and acoustic characteristics. Additionally, the words are tokenized using the BERT tokenizer [31].

TABLE 4. OpenFace Feature Index

Group	Index (i)	Feature	Length
Eye Gaze	0-2	Leftmost Eye Gaze	3
	3-5	Rightmost Eye Gaze	3
	6-7	Eye Gaze Direction	2
	8-119	Eye Landmark 2D	112
	120-287	Eye Landmark 3D	168
Head Pose	288-290	Head Location	3
	291-293	Rotation	3
Facial Landmark	294-429	Landmark 2D	136
	430-633	Landmark 3D	204
Face Shape	634-639	Rigid Face Shape	6
	640-673	Non-rigid Face Shape	34
Facial Action Unit	674-708	Facial Action Unit	35
Total			**709**

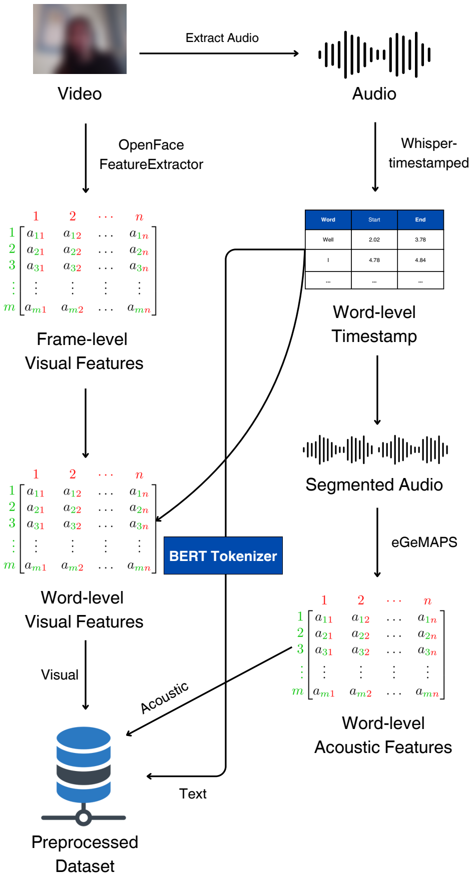


FIGURE 3. Dataset Preprocessing Outline

Subsequently, the timestamps are employed to partition the audio recording into segments delineated by word boundaries, thereby generating distinct snippets for each uttered word. Within each snippet, acoustic features are extracted using eGeMAPS [32], capturing attributes such as pitch, energy, and speech rate. This process yields a feature vector comprising 88 dimensions for each spoken word.

Subsequently, OpenFace [33] is employed to capture visual characteristics such as posture and gestures. Unlike audio processing, the entire video is utilized to extract visual features, resulting in a feature vector of 709 dimensions. The feature vector for each spoken word is obtained by locating the frame that corresponds to each word in the visual features.

After extracting and aligning features from each modality, they are consolidated into a cohesive dataset. Each entry in this dataset encapsulates a succinct record consisting of three elements: a triplet of textual, acoustic, and visual features (*L*, *A*, *V*), the corresponding label for each data point *y*, and the segment ID. This procedure, as shown in Fig. 3, is iterated across the training, validation, and testing datasets.

D. HYPERPARAMETER TUNING

In this work, the MAG-BERT-ARL framework, comprising MAG-BERT and Adversarially Reweighted Learning (ARL), was adopted. Existing implementations of MAG-BERT and ARL can be found [18], [19]. However, the reproduced PyTorch implementation of ARL was used to maintain consistency, as MAG-BERT is also implemented in PyTorch [46].

The integration process began with replacing the linear MLP head in MAG-BERT with ARL for sequence classification. This was the only modification required on the MAG-BERT side. For the ARL component, skip connections and batch normalization were implemented. This involved modifying the hidden layers in both the learner and the adversary by adding a dropout layer ($p = 0.5$) and a layer norm after the linear layer. Additionally, all hidden layer sizes were adjusted to match the input or embedding size to ensure compatibility with the CLS token from MAG-BERT via skip connections. A post-norm configuration of batch normalization was applied by passing the input through the linear and dropout layers, subsequently combining it with the unmodified input before applying layer norm. This was implemented in all layers except the first. Finally, the sigmoid function in the adversary was replaced with ReLU.

For MAG-BERT, the optimal configuration outlined in the original paper was adhered to. For ARL, the learner was constructed as a three-layer feed-forward network, while the adversary was formulated as a one-layer feed-forward network with ReLU activation functions. The remaining training hyperparameters are presented in Table 5.

During training on the ETS dataset, both training and validation batch sizes were set to $m = 64$ and the maximum sequence length to $q = 256$. For the FI dataset, both training and validation batch sizes were set to $m = 128$ and the maximum sequence length to $q = 176$. The model was trained on all datasets for 200 epochs without early stopping. However, the baseline model employed early stopping, with a patience parameter set to 2. The AdamW optimizer was used for training, with learning rates of $\eta_{learner} = 10^{-5}$ and $\eta_{adversary} = 10^{-5}$. The optimizer was separated for the learner and the classifier.

E. EVALUATION METRICS

Metrics for both performance and fairness are considered in the evaluation. For performance, both classification and regression metrics are assessed. Regression metrics are calculated based on the model's predictions and the ground truth hiring scores. For classification metrics, a decision threshold specific to each dataset is applied to convert continuous values into binary classification labels.

1) Classification Evaluation

Accuracy, precision, recall, and F1 score are chosen for the classification task. Higher accuracy, precision, recall, and F1 score signify better model performance in terms of accurately predicting the interviewee's suitability for the job.

For positive class predictions, the right prediction is called true positive (*TP*), while the wrong one, that the truth was actually negative class, is called false positive (*FP*). For negative class predictions, the right prediction is called true negative (*TN*), while the wrong one, that the truth was actually positive class, is called false negative (*FN*). These terms are important since they are used a lot in classification metrics.

Accuracy is the percentage of correct classifications. If accuracy is 1, it implies that the model prediction always predict right. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

Precision or confidence is the percentage of correct classification among the positive class predictions. If precision is 1, it implies that the model always predict right when it comes to predicting the positive class. Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall or sensitivity is the percentage of correct classification among the actual positive instances. If recall is 1, it implies that the model always catch up on actual positive instances. Recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

F-score is harmonic mean between precision and recall. F-score is defined as follows:

$$\text{F-score} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \tag{12}$$

In this equation, β is a hyperparameter. Higher β implies more importance to recall. When $\beta = 1$, precision and recall is of the same importance. In this case, F-score is also called F1 score.

2) Regression Evaluation

Furthermore, the Pearson correlation coefficient r_p and root mean squared error (RMSE) between the predicted hiring scores and the actual hiring scores are calculated for the regression task. A high positive Pearson correlation coefficient r_p value indicates that the model's predictions tend to align with the human judgments, while a low RMSE value signifies that the model's predictions are close to the actual scores.

Pearson correlation coefficient (r_p) is a parametric measure that assesses the linear relationship between two random

TABLE 5. Hyperparameters for Training

Hyperparameter	Value for ETS Dataset	Value for FI Dataset
`Max Seq Length`	256	176
`Train Batch Size`	64	128
`Num Labels`	1	1
`Dev Batch Size`	64	128
`N Epochs`	200	200
`Beta Shift`	1.0	1.0
`Dropout Prob`	0.5	0.5
`Model`	`bert-base-uncased`	`bert-base-uncased`
`Tokenizer`	`bert-base-uncased`	`bert-base-uncased`
`Learning Rate`	10^{-5}	10^{-5}
`Gradient Accumulation Step`	1	1
`Warmup Proportion`	0.1	0.1
`Seed`	8	8
`Learning Rate Adversary`	10^{-5}	10^{-5}
`Pretrain Steps`	6	6

variables [47]. It ranges from +1, indicating perfect positive correlation, to -1, indicating perfect negative correlation, with 0 representing no linear relationship. Pearson correlation is defined as follows:

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{13}$$

In this context, x and y represent the variables for which the correlation is to be calculated, while $\overline{x}$ and $\overline{y}$ denote the respective means of x and y.

Root mean squared error (RMSE) is the square root of the mean squared difference between predictions and true values. If RMSE is 0, it implies that the predictions and true values match exactly. RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x - y)^2} \tag{14}$$

3) Fairness Evaluation

For fairness metrics, demographic parity (DP) (as in Eq, 15), equalized odds (EO) (as in Eq, 16), and equal accuracy (EA) (as in Eq, 17) metrics are utilized to assess fairness in the model's predictions across different demographic groups. This ensures that the model's predictions are independent of sensitive attributes such as race, gender, or age.

Demographic parity evaluates whether the proportion of positive outcomes (e.g., job offers) is the same across different demographic groups [48]. Demographic parity for a model with prediction $\hat{y}$ and sensitive attribute A is defined as follows:

$$\text{DP} = P(\hat{y} = 1 | A = 0) - P(\hat{y} = 1 | A = 1) \tag{15}$$

If demographic parity holds, that is it is equal to zero, it implies that the model's predictions are independent of the sensitive attribute.

Equalized odds assesses whether the true positive rates (sensitivity/recall) and false positive rates are the same across different demographic groups [49]. Equalized odds for a model with prediction $\hat{y}$, true value y, and sensitive attribute A is defined as follows:

$$\text{EO} = P(\hat{y} = 1 | y = 1, A = 0) - P(\hat{y} = 1 | y = 1, A = 1) \tag{16}$$

Unlike demographic parity, equalized odds assess whether or not the model's predictions are equally accurate across different demographic groups.

Equal Accuracy (EA) ensures that the model performs equally well across different demographic groups [11]. This means that the estimation error, as measured by a chosen metric, M, should be similar for all groups. Equal accuracy for a model with prediction $\hat{y}$, true value y, and sensitive attribute A is defined as follows:

$$\text{EA} = M(y, \hat{y} | A = 0) - M(y, \hat{y} | A = 1) \tag{17}$$

If the Equal Accuracy (EA) is zero, it implies that the model's prediction errors are similar across all demographic groups.

Demographic parity requires that the distribution of predicted outcomes is consistent across different demographic groups, while equalized odds additionally demands that the true positive rates (sensitivity) and false positive rates (specificity) are equal across these groups. Equal accuracy demands that the overall performance of the model is consistent across different demographic groups. Note that the sensitive attribute is only utilized during testing to facilitate the computation of these metrics.

F. GENDER CLASSIFICATION

To assess group fairness between genders in the ETS dataset, where gender annotations are absent, an automatic gender classification model was employed to annotate the dataset with gender labels. A fine-tuned Vision Transformer (ViT) model from HuggingFace was used to classify each video as either male or female [50]. The model takes an image as input and subsequently classifies it as either male or female. For classifying video interviews in the ETS dataset, one to three random frames were extracted from each video using OpenCV [51] to serve as the image representation for that particular video. These frames were subsequently classified using the aforementioned model. Finally, a voting algorithm was utilized to determine the video's gender classification based on the majority label of the frames. It should be noted that these labels were only utilized for the evaluation of group fairness metrics within the ETS dataset.

G. INTERPRETATION FRAMEWORK

Similar framework utilized in HirePreter [30] was utilized to interpret MAG-BERT-ARL. However, Gradient SHAP was applied to both textual and non-textual features, whereas HirePreter's framework utilized Gradient SHAP [23] for non-textual feature and integrated gradients [34] for textual feature. The Captum library's implementation of the Gradient SHAP primary attribution algorithm was utilized to calculate Gradient SHAP [52]. The baseline for calculating Gradient SHAP $\hat{x}$ was set to a zero matrix for both acoustic and visual features. For the text features, the baseline was defined as $\hat{x} = w + \alpha \times (x - w)$, where w and α are matrices filled with random numbers in the interval [0, 1).

V. RESULTS AND ANALYSES

This section presents the model's performance in predicting hiring decisions, fairness related to hiring decisions, impact of nonverbal features, and insights into factors influencing the model's predictions.

A. OVERALL EVALUATION

Table 6 summarizes the ranking results across various models for fairness, regression, and classification tasks assessed on ETS and FI dataset. Among the models evaluated, MAG-BERT-ARL M emerges as a standout performer on the ETS dataset. This can be attributed to its regression and classification performance. However, its counterpart, MAG-BERT-ARL B and MAG-BERT-ARL MB, while exhibiting sufficient performance, falls slightly short in comparison, particularly in terms of regression and classification ranking compared to MAG-BERT-ARL M on the ETS dataset. MAG-BERT shares the top spot with MAG-BERT-ARL MB in fairness ranking on the ETS dataset. Notably, all MAG-BERT-ARL variants ranks higher than the baseline MAG-BERT model.

On the FI dataset, MAG-BERT-ARL MB emerges as a standout performer. This can be attributed to its fairness and classification performance. Conversely, MAG-BERT-ARL M and MAG-BERT-ARL B secure slightly lower ranks, primarily due to their reduced fairness and classification scores. Notably, all MAG-BERT-ARL variants perform better than the baseline MAG-BERT model.

While the overall performance of MAG-BERT-ARL is better than that of MAG-BERT, it is notable that the top-ranking variant differs between the two datasets. For the ETS dataset, MAG-BERT-ARL M is the top performer, while for the FI dataset, it is MAG-BERT-ARL MB. Fig. 4 depicts the prediction and dataset distribution of all MAG-BERT-ARL variants overlaid on the ETS dataset and FI datasets. The figure suggests that adding binary cross-entropy loss increases the variance in the model's predictions. It can also be observed that the distribution of the ETS dataset is more skewed compared to the distribution of the FI dataset. Consequently, more of MAG-BERT-ARL MB's predictions for the ETS dataset lie outside its distribution. In contrast, for the FI dataset, MAG-BERT-ARL MB's predictions cover a broader range of the FI dataset, reducing errors that were previously concentrated around the mean. Therefore, MAG-BERT-ARL MB performs better on the FI dataset, while MAG-BERT-ARL M performs better on the ETS dataset.

B. HIRING DECISION PREDICTION

The Pearson correlation coefficient r_p evaluation presented in Table 7 shows that all MAG-BERT-ARL variants improve r_p. MAG-BERT-ARL B achieves the best r_p evaluation for both ETS and FI datasets. The model increases the r_p compared to the baseline MAG-BERT model by 0.05 and 0.17 for both ETS and FI datasets.

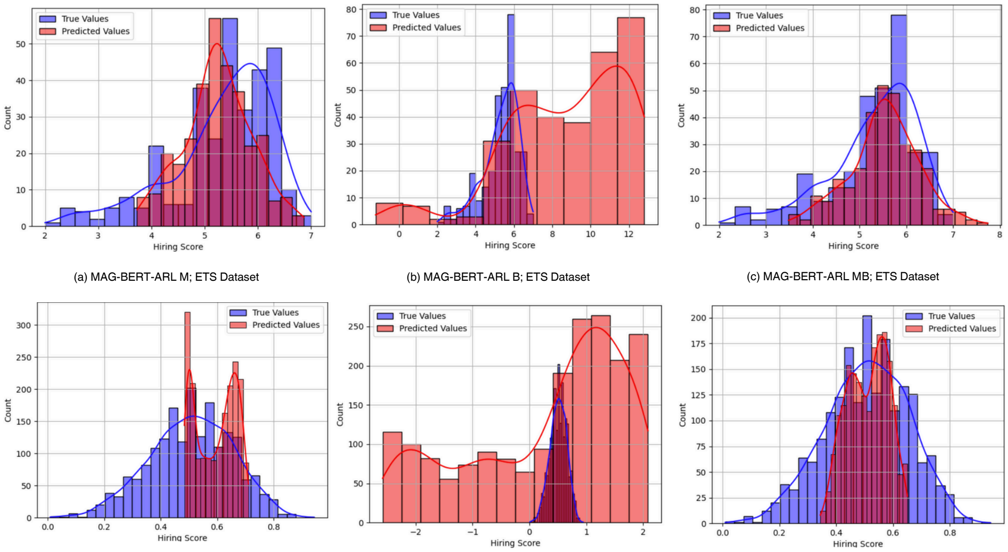
Root mean squared error (RMSE) evaluation shows MAG-BERT-ARL models provide sufficient RMSE results. MAG-BERT-ARL B model has worst compared to any model for both ETS and FI datasets. MAG-BERT-ARL M achieves the best RMSE evaluation for ETS dataset, improving the RMSE by 0.12 compared to the baseline MAG-BERT model. Different results are found in FI dataset. In this case, RMSE for MAG-BERT-ARL M (0.16) is worse than MAG-BERT-ARL MB (0.14) which gives the best RMSE evaluation. MAG-BERT-ARL MB achieves the best RMSE evaluation for FI dataset, improving the RMSE by 0.05 compared to the baseline MAG-BERT model.

The overall performance of the models is calculated using the rank (Rank column) of the regression ranks average (Avg column) presented in Table 7. For both ETS and FI datasets, all MAG-BERT-ARL variants rank higher than the baseline MAG-BERT model. For ETS dataset, the best model is MAG-BERT-ARL M with a regression ranks average value of 1.5, followed by model MAG-BERT-ARL B and MAG-BERT-ARL MB. For FI dataset, the best model is MAG-BERT-ARL MB with a regression ranks average value of 1.75, followed by model MAG-BERT-ARL M. The best model MAG-BERT-ARL M and MAG-BERT-ARL MB are both a MAG-BERT-ARL model that apply MSE loss function and MSE combined with binary cross entropy loss function, respectively. There is no reported regression performance for the [17] and [11] model, which is marked with dashes.

The accuracy (Acc) evaluation presented in Table 8 shows that all MAG-BERT-ARL variants, except MAG-BERT-ARL B for the ETS dataset, improve accuracy. The MAG-BERT-ARL M variant achieves the best accuracy evaluation for the ETS dataset, improving accuracy by 0.08 compared to the baseline MAG-BERT model. Conversely, for the FI dataset, the accuracy for the MAG-BERT-ARL M variant (0.58) is

TABLE 6. Ranking results for fairness, regression, and classification. Fairness rank is the rank of the average rank of the model fairness evaluation score, Regression rank is the rank of the average rank of the model regression evaluation score, Classification rank is the rank of the average rank of the model classification evaluation score, Overall rank is the rank of the average rank of Fairness rank, Regression rank, and Classification rank. MAG-BERT-ARL M is a MAG-BERT-ARL model that applies MSE loss function, MAG-BERT-ARL B is a MAG-BERT-ARL model that applies binary cross-entropy loss function, and MAG-BERT-ARL MB is a MAG-BERT-ARL model that applies MSE combined with binary cross entropy loss function.

Model	Fairness Rank ↓	Regression Rank ↓	Classification Rank ↓	Overall Rank ↓
		ETS Dataset		
MAG-BERT	**1.5**	4.0	4.0	4.0
MAG-BERT-ARL M	2.0	**1.0**	**1.0**	**1.0**
MAG-BERT-ARL B	3.0	3.0	2.5	3.0
MAG-BERT-ARL MB	**1.5**	2.0	2.5	2.0
		FI Dataset		
MAG-BERT	4.0	4.0	4.0	4.0
MAG-BERT-ARL M	2.5	3.0	2.0	3.0
MAG-BERT-ARL B	2.5	**1.5**	3.0	2.0
MAG-BERT-ARL MB	**1.0**	**1.5**	**1.0**	**1.0**



(a) MAG-BERT-ARL M; ETS Dataset

(b) MAG-BERT-ARL B; ETS Dataset

(c) MAG-BERT-ARL MB; ETS Dataset

(d) MAG-BERT-ARL M; FI Dataset

(e) MAG-BERT-ARL B; FI Dataset

(f) MAG-BERT-ARL MB; FI Dataset

FIGURE 4. Distribution of MAG-BERT-ARL prediction laid on ETS and FI Datasets. The blue-colored distribution represents the actual values, while the red-colored distribution represents the model's predictions. The top distribution represents the ETS dataset, while the bottom distribution corresponds to the FI dataset. From left to right, the distributions are shown for MAG-BERT-ARL M, MAG-BERT-ARL B, and MAG-BERT-ARL MB.

worse than that of the MAG-BERT-ARL B and MAG-BERT-ARL MB variants (0.66), which provide the best accuracy evaluation. The MAG-BERT-ARL B and MAG-BERT-ARL MB variants achieve the best accuracy evaluation for the FI dataset, improving accuracy by 0.17 compared to the baseline MAG-BERT model. However, any of the MAG-BERT-ARL variants demonstrate inferior performance compared to the models presented by [11], [30], and [16], although there are slight differences in [30]'s and [16]'s datasets.

The precision (Prec) evaluation shows that all MAG-BERT-ARL variants outperform the baseline MAG-BERT model. The MAG-BERT-ARL B variant achieves the best precision evaluation for the ETS dataset, increasing precision by 0.39 compared to the baseline MAG-BERT model. The MAG-BERT-ARL B and MAG-BERT-ARL M variants show higher precision evaluation (0.72 and 0.66, respectively) compared to the model presented by [17] (0.65). For the FI dataset, the MAG-BERT-ARL B and MAG-BERT-ARL MB variants achieve the best precision evaluation, increasing precision by 0.06 compared to the baseline MAG-BERT model.

The recall (Rec) evaluation shows that all MAG-BERT-ARL variants, except MAG-BERT-ARL B for the ETS dataset, improve the F1 score. The MAG-BERT-ARL M variant achieves the best recall evaluation for the ETS dataset, increasing it by 0.08 compared to the baseline MAG-BERT model. The MAG-BERT-ARL M variant also shows higher recall evaluation (0.66) compared to the model presented by [17] (0.65). For the FI dataset, the MAG-BERT-ARL B and MAG-BERT-ARL MB variants achieve the best recall evaluation, increasing recall by 0.17 compared to the baseline MAG-BERT model.

The evaluation of the F1 score demonstrates that all variants of MAG-BERT-ARL surpass the baseline MAG-BERT model. Notably, MAG-BERT-ARL M exhibits the highest F1 score improvement in the ETS dataset, increasing the F1

TABLE 7. Regression results for ETS and FI datasets. r_p is the Pearson correlation coefficient (higher scores are better), RMSE is root mean squared error (lower scores are better), Avg is the average rank of the model regression evaluation score, Rank is the rank of Avg. MAG-BERT-ARL M is a MAG-BERT-ARL model that applies MSE loss function, MAG-BERT-ARL B is a MAG-BERT-ARL model that applies binary cross-entropy loss function, and MAG-BERT-ARL MB is a MAG-BERT-ARL model that applies MSE combined with binary cross entropy loss function.

Model	Metrics		Regression Rank			
	r_p ↑	RMSE ↓	r_p ↓	RMSE ↓	Avg ↓	Rank ↓
			ETS Dataset			
ETS Dataset	1.00	0.00				
Chen et al. (2017)	-	-				
MAG-BERT	0.43	0.95	4.0	3.0	3.50	4.0
MAG-BERT-ARL M	0.53	**0.83**	2.0	**1.0**	**1.50**	**1.0**
MAG-BERT-ARL B	**0.59**	4.38	**1.0**	4.0	2.50	2.5
MAG-BERT-ARL MB	0.48	0.91	3.0	2.0	2.50	2.5
			FI Dataset			
FI Dataset	1.00	0.00				
Yan et al. (2020)	-	-				
MAG-BERT	0.23	0.19	4.0	3.0	3.50	4.0
MAG-BERT-ARL M	0.38	0.16	2.5	2.0	2.25	2.0
MAG-BERT-ARL B	**0.40**	1.33	**1.0**	4.0	2.50	3.0
MAG-BERT-ARL MB	0.38	**0.14**	2.5	**1.0**	**1.75**	**1.0**

TABLE 8. Classification results for ETS and FI datasets. Acc is accuracy (higher scores are better), Prec is precision (higher scores are better), Rec is recall (higher scores are better), F1 is F1 score (higher scores are better), Avg is the average rank of the model classification evaluation score, Rank is the rank of Avg.

Model	Metrics				Classification Rank					
	Acc ↑	Prec ↑	Rec ↑	F1 ↑	Acc ↓	Prec ↓	Rec ↓	F1 ↓	Avg ↓	Rank ↓
					ETS Dataset					
ETS Dataset	1.00	1.00	1.00	1.00						
Chen et al. (2017)	-	0.65	0.65	0.65						
MAG-BERT	0.58	0.33	0.58	0.42	3.0	4.0	3.0	4.0	3.50	4.0
MAG-BERT-ARL M	**0.66**	0.66	**0.66**	**0.66**	**1.0**	2.0	**1.0**	**1.0**	**1.25**	**1.0**
MAG-BERT-ARL B	0.53	**0.72**	0.53	0.61	4.0	**1.0**	4.0	3.0	3.00	3.0
MAG-BERT-ARL MB	0.63	0.64	0.63	0.64	2.0	3.0	2.0	2.0	2.25	2.0
					FI Dataset					
FI Dataset	1.00	1.00	1.00	1.00						
Yan et al. (2020)	**0.92**	-	-	-						
MAG-BERT	0.49	0.60	0.49	0.54	4.0	4.0	4.0	4.0	4.00	4.0
MAG-BERT-ARL M	0.58	0.62	0.58	0.60	3.0	3.0	3.0	3.0	3.00	3.0
MAG-BERT-ARL B	0.66	**0.66**	**0.66**	**0.66**	**1.5**	**1.5**	**1.5**	**1.5**	**1.50**	**1.5**
MAG-BERT-ARL MB	0.66	**0.66**	**0.66**	**0.66**	**1.5**	**1.5**	**1.5**	**1.5**	**1.50**	**1.5**

score by 0.24 compared to the baseline and 0.01 compared to the Chen et al. (2017) model. In the FI dataset, both MAG-BERT-ARL B and MAG-BERT-ARL MB variants achieve superior F1 score, increasing it by 0.12 relative to the baseline MAG-BERT model.

The overall performance of the models is calculated using the rank (Rank column) of the classification ranks average (Avg column) presented in Table 8. For both the ETS and FI datasets, all MAG-BERT-ARL variants rank higher than the baseline MAG-BERT model. For the ETS dataset, the best model is the MAG-BERT-ARL M variant, with a classification ranks average value of 1, followed by the MAG-BERT-ARL B variant. For the FI dataset, the best models are the MAG-BERT-ARL B and MAG-BERT-ARL MB variants, with a classification ranks average value of 1.5, followed by the MAG-BERT-ARL M variant. The best models, MAG-BERT-ARL M, MAG-BERT-ARL B, and MAG-BERT-ARL MB, apply the MSE loss function, binary cross-entropy loss function, and the MSE combined with binary cross-entropy loss function, respectively.

C. GROUP FAIRNESS

The demographic parity (DP) evaluation presented in Table 9 shows that all MAG-BERT-ARL variants falls short of matching the DP evalution of the baseline MAG-BERT model in ETS dataset. Conversely, all MAG-BERT-ARL variants shows lower DP evaluation compared to the baseline MAG-BERT model in the FI dataset. For the ETS dataset, the MAG-BERT-ARL MB variant achieves a lower DP evaluation than ETS dataset and out of all MAG-BERT-ARL variants, decreasing DP by 0.017 compared to the DP evaluation of the ETS dataset. For the FI dataset, the MAG-BERT-ARL MB variant achieves the best DP evaluation for the ETS dataset, decreasing it by 0.016 compared to the baseline MAG-BERT model. The MAG-BERT-ARL MB variant also shows lower DP evaluation (0.001) compared to the DP

TABLE 9. **Fairness performance results for ETS and FI datasets. DP is demographic parity (lower scores are better), EO is equalized odds (lower scores are better), EA is equal accuracy (lower scores are better), Avg is the average rank of the model fairness evaluation score, Rank is the rank of Avg.**

Model	Metrics			Fairness Rank				
	DP ↓	EO ↓	EA ↓	DP ↓	EO ↓	EA ↓	Avg ↓	Rank ↓
ETS Dataset								
ETS Dataset	0.029	0.000	0.000					
Chen et al. (2017)	-	-	-					
MAG-BERT	**0.000**	**0.000**	0.131	**1.0**	**1.0**	3.0	**1.67**	**1.5**
MAG-BERT-ARL M	0.072	0.117	0.037	3.0	3.0	2.0	2.67	2.0
MAG-BERT-ARL B	0.080	0.137	0.197	4.0	4.0	4.0	4.00	3.0
MAG-BERT-ARL MB	0.012	0.063	**0.019**	2.0	2.0	**1.0**	**1.67**	**1.5**
FI Dataset								
FI Dataset	0.011	0.000	0.000					
Yan et al 2020	0.002	-	**0.001**					
MAG-BERT	0.017	**0.022**	0.032	4.0	**1.0**	4.0	3.00	4.0
MAG-BERT-ARL M	0.006	0.024	0.010	3.0	2.0	3.0	2.67	2.5
MAG-BERT-ARL B	0.002	0.085	0.007	2.0	4.0	2.0	2.67	2.5
MAG-BERT-ARL MB	**0.001**	0.036	0.004	**1.0**	3.0	**1.0**	**1.67**	**1.0**

evaluation of FI dataset (0.011) and model presented by [11] (0.002).

The equalized odds (EO) evaluation shows no improvement for both ETS and FI datasets. All MAG-BERT-ARL variants shows higher EO evaluation compared to the baseline MAG-BERT model in both ETS and FI datasets. There is no reported EO evaluation for the [11] model, which is marked with dashes.

The evaluation of the equal accuracy (EA) demonstrates that all variants of MAG-BERT-ARL, except MAG-BERT-ARL B for the ETS dataset, surpass the baseline MAG-BERT model. The MAG-BERT-ARL MB variant provides the lowest EA evaluation for both ETS and FI datasets. The MAG-BERT-ARL MB variant decreases the DP compared to the baseline MAG-BERT model by 0.112 and 0.028 for both ETS and FI datasets.

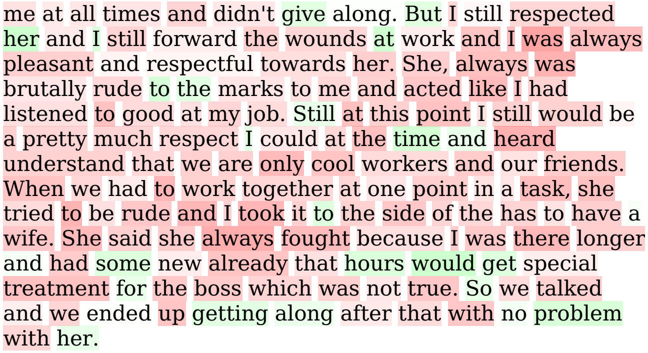
The overall performance of the models is calculated using the rank (Rank column) of the fairness ranks average (Avg column) presented in Table 9. For both the ETS and FI datasets, all MAG-BERT-ARL variants rank comparable to or higher than the baseline MAG-BERT model. For the ETS dataset, MAG-BERT-ARL MB ties for first place in the overall fairness ranking alongside the baseline MAG-BERT model, with a fairness ranks average value of 1.5, followed by the MAG-BERT-ARL M variant. For the FI dataset, the best model is the MAG-BERT-ARL MB variant, with a fairness ranks average value of 1, followed by the MAG-BERT-ARL M and MAG-BERT-ARL B variants. The best models, MAG-BERT-ARL M and MAG-BERT-ARL B, apply the MSE loss function and binary cross-entropy loss function, respectively.

D. ABLATION TEST ON MULTIMODALITIES

To evaluate the importance of each modality within the model, an ablation test was performed. The evaluation of fairness, classification, and regression metrics was conducted for four distinct combinations of three modalities: text (T), text and acoustic (T+A), text and visual (T+V), and text, acoustic, and visual (T+A+V). These combinations were assessed on the ETS dataset [17] using the MAG-BERT-ARL model.

Findings from the ETS dataset in Table 10 indicate that MAG-BERT-ARL with only the text modality (T) performed the highest in fairness metrics. The addition of the acoustic modality (A) leads to a significant improvement in classification evaluation metrics. However, it results in a decline in both fairness and regression evaluation metrics. Similarly, incorporating the visual modality (V) yields comparable effects, with the added benefit of enhanced regression evaluation metrics. This suggests a bias-accuracy trade-off when additional modalities, such as acoustic (A) and visual (V), are included. However, the highest accuracy, recall, F1 score, and RMSE were achieved by MAG-BERT-ARL when all three modalities were integrated. This underscores the idea that incorporating multiple modalities substantially improves the model's classification and regression performance on the ETS dataset.

E. INTERPRETATION



me at all times and didn't give along. But I still respected her and I still forward the wounds at work and I was always pleasant and respectful towards her. She, always was brutally rude to the marks to me and acted like I had listened to good at my job. Still at this point I still would be a pretty much respect I could at the time and heard understand that we are only cool workers and our friends. When we had to work together at one point in a task, she tried to be rude and I took it to the side of the has to have a wife. She said she always fought because I was there longer and had some new already that hours would get special treatment for the boss which was not true. So we talked and we ended up getting along after that with no problem with her.

FIGURE 5. **Word importance calculated from a random instance's transcribed answer in the ETS dataset to the prompt: "Please tell us about a time when you had to work with someone you did not especially like or get along with. How did you interact?". Words highlighted in green positively influence the hiring recommendation score, whereas those highlighted in red negatively affect the score.**

TABLE 10. MAG-BERT-ARL fairness, classification, and regression results on different modalities for ETS dataset. T stands for the text modality, A for the acoustic modality, and V for the visual modality.

Method	Fairness			Classification				Regression	
	DP ↓	EO ↓	EA ↓	Acc ↑	Prec ↑	Rec ↑	F1 ↑	r_p ↑	RMSE ↓
ETS Dataset	0.029	0.000	0.000	1.00	1.00	1.00	1.00	1.00	0.00
T	**0.046**	**0.076**	**0.018**	0.50	**0.77**	0.50	0.60	**0.63**	1.64
T+A	0.084	0.109	0.129	0.62	0.63	0.62	0.63	0.41	1.71
T+V	0.080	0.181	0.105	0.60	0.62	0.60	0.61	0.38	1.40
T+A+V	0.072	0.117	0.037	**0.66**	0.66	**0.66**	**0.66**	0.53	**0.83**

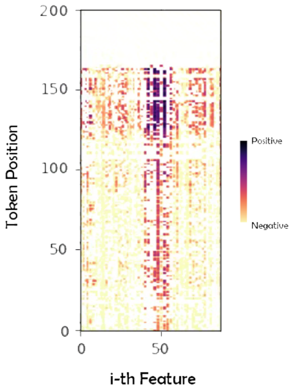


FIGURE 6. Gradient SHAP attribution scores for acoustic features. The x-axis represents the feature index for the eGeMAPS acoustic feature set, while the y-axis indicates the token position where the scores are calculated. Dark purple signifies a positive contribution to the hiring recommendation score, whereas light yellow indicates a negative contribution.

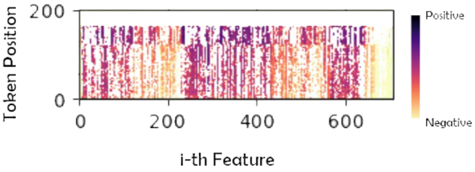


FIGURE 7. Gradient SHAP attribution scores for visual features. The x-axis represents the feature index for the OpenFace acoustic feature set, while the y-axis indicates the token position where the scores are calculated. Dark purple signifies a positive contribution to the hiring recommendation score, whereas light yellow indicates a negative contribution.

To evaluate the impact of each token within each modality on the overall hiring decision, Gradient SHAP was calculated. An instance was randomly selected from the ETS dataset for this purpose. The modalities for text, acoustic, and visual features were analyzed using Gradient SHAP to interpret the model's prediction for this specific instance.

Fig. 5 shows the textual interpretation, where the word importance of the text is illustrated. For word importance, the color green indicates that the word is likely to increase the probability of being hired, whereas the color red suggests the opposite effect. In this particular example, words such as “give”, “get”, and “getting along” are associated with an increased hiring recommendation score, as positive attributes such as professionalism, interpersonal skills, and conflict resolution are reflected. Conversely, words like “rude”, “brutally”, “wounds”, and “fought” may carry negative connotations, which adversely affect the score.

In Fig. 6 and Fig. 7, the feature importance for acoustic and visual features extracted using eGeMAPS [32] and OpenFace [33], respectively, is shown. For feature importance, the color dark purple represents that a particular feature is likely to increase the hiring probability, and light yellow signifies the opposite. It has been revealed through analysis that the attribution scores are notably higher towards the end of the video, indicating that hiring decisions are significantly influenced by the video's conclusion. Moreover, specific features were found to have a greater impact on hiring decisions. Notably, among acoustic features, the vocal tract attributes (including F1-3 frequency, bandwidth, and amplitude) were identified as the most influential by the model, whereas action units within visual features were deemed less significant. This observation may partly account for the slight increase in Equalized Odds (EO), as variations in vocal tract characteristics across age and gender [53] could contribute to differences in hiring decisions.

VI. CONCLUSION

To address fairness without sensitive attributes, an automated interview assessment system was developed, prioritizing fairness by excluding sensitive attributes such as gender, race, and age during the training and validation phases. MAG-BERT-ARL, which combines MAG-BERT [18] with Adversarially Reweighted Learning (ARL) [19], was proposed for achieving this goal. Three variants of MAG-BERT-ARL were developed based on their loss function to observe which models can improve the baseline model performance in accuracy and fairness. Additionally, an ablation study and model interpretations for each model modality is presented.

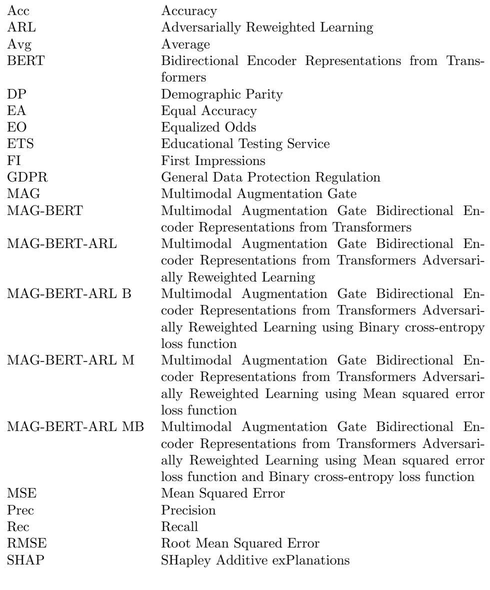
The findings demonstrate that integrating ARL, a technique for achieving fairness without demographics, to MAG-BERT can enhance both the fairness and performance of a multimodal system. The enhancements range from 0.05 to 0.17 for the Pearson correlation coefficient, 0.05 to 0.12 for root mean squared error, 0.06 to 0.39 for accuracy, precision, recall, and F1 score, and 0.028 to 0.112 for equal accuracy. Despite these enhancements, it was noted that MAG-BERT-ARL primarily addresses fairness by enhancing accuracy for underrepresented groups, as evidenced by the enhancements in equal accuracy.

Several limitations may have influenced the results and findings. First, this research focused exclusively on gender as the demographics. In terms of fairness evaluations, it was observed that while equal accuracy metrics consistently improved for both the Educational Testing Service (ETS) [17] and the First Impressions (FI) [24] datasets, demographic parity metrics exhibited inconsistent improvements, with equalized odds showing worsened performance. Furthermore, despite the model demonstrating the ability for fairness without demographics, as well as observability, which aligns with regulatory compliance standards, it should be noted that formal validation for compliance has not been conducted.

Future research will focus on exploring alternative techniques for achieving fairness without demographics to further enhance fairness. Investigations will include integrating methods such as exploring alternative modalities or feature sets and employing fair representation learning to debias modalities that may carry inherent biases. Furthermore, efforts will be directed towards incorporating fairness without demographics techniques into better-performing multimodal foundation models to achieve prediction accuracy comparable to state-of-the-art models.

.

APPENDIX A KEY TERMS



Acc	Accuracy
ARL	Adversarially Reweighted Learning
Avg	Average
BERT	Bidirectional Encoder Representations from Transformers
DP	Demographic Parity
EA	Equal Accuracy
EO	Equalized Odds
ETS	Educational Testing Service
FI	First Impressions
GDPR	General Data Protection Regulation
MAG	Multimodal Augmentation Gate
MAG-BERT	Multimodal Augmentation Gate Bidirectional Encoder Representations from Transformers
MAG-BERT-ARL	Multimodal Augmentation Gate Bidirectional Encoder Representations from Transformers Adversarially Reweighted Learning
MAG-BERT-ARL B	Multimodal Augmentation Gate Bidirectional Encoder Representations from Transformers Adversarially Reweighted Learning using Binary cross-entropy loss function
MAG-BERT-ARL M	Multimodal Augmentation Gate Bidirectional Encoder Representations from Transformers Adversarially Reweighted Learning using Mean squared error loss function
MAG-BERT-ARL MB	Multimodal Augmentation Gate Bidirectional Encoder Representations from Transformers Adversarially Reweighted Learning using Mean squared error loss function and Binary cross-entropy loss function
MSE	Mean Squared Error
Prec	Precision
Rec	Recall
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations

FIGURE 8. **Key Terms and Their Meanings**

ACKNOWLEDGMENT

Part of this research is funded by Faculty of Computer Science, Universitas Indonesia, under Grant NKB-7/UN2.F11.D/HKP.05.00/2024 and supported by Japan Science and Technology Agency (JST) Sakura Science Program.

REFERENCES

[1] C. Fernández-Martínez and A. Fernández, "AI and recruiting software: Ethical and legal implications," Paladyn J. Behav. Robotics, vol. 11, no. 1, pp. 199–216, 2020. [Online]. Available: https://doi.org/10.1515/pjbr-2020-0030

[2] E.-R. Lukacik, J. S. Bourdage, and N. Roulin, "Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews," Human Resource Management Review, vol. 32, no. 1, p. 100789, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053482220300620

[3] HireVue, "About the company: Leadership & ceo: Hirevue," accessed: 2024-05-29. [Online]. Available: https://www.hirevue.com/about

[4] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," 2020. [Online]. Available: https://doi.org/10.3389/frai.2020.561802

[5] A. Siapka, "The ethical and legal challenges of artificial intelligence: The eu response to biased and discriminatory ai," Available at SSRN 3408773, 2018.

[6] L. Antunes, M. Naldi, G. F. Italiano, K. Rannenberg, and P. Drogkaris, Eds., Privacy Technologies and Policy - 8th Annual Privacy Forum, APF 2020, Lisbon, Portugal, October 22-23, 2020, Proceedings, ser. Lecture Notes in Computer Science. Springer, 2020, vol. 12121. [Online]. Available: https://doi.org/10.1007/978-3-030-55196-4

[7] I. Ajunwa, "An auditing imperative for automated hiring systems," Harv. JL & Tech., vol. 34, p. 621, 2020.

[8] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, "What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness," in FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, M. C. Elish, W. Isaac, and R. S. Zemel, Eds. ACM, 2021, pp. 249–260. [Online]. Available: https://doi.org/10.1145/3442188.3445888

[9] M. van Bekkum and F. J. Z. Borgesius, "Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?" Comput. Law Secur. Rev., vol. 48, p. 105770, 2023. [Online]. Available: https://doi.org/10.1016/j.clsr.2022.105770

[10] E. Derman, "Dataset bias mitigation through analysis of CNN training scores," CoRR, vol. abs/2106.14829, 2021. [Online]. Available: https://arxiv.org/abs/2106.14829

[11] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in ICMI '20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020, K. P. Truong, D. Heylen, M. Czerwinski, N. Berthouze, M. Chetouani, and M. Nakano, Eds. ACM, 2020, pp. 361–369. [Online]. Available: https://doi.org/10.1145/3382507.3418889

[12] J. Wang, Y. Liu, and X. E. Wang, "Are gender-neutral queries really gender-neutral? mitigating gender bias in image search," CoRR, vol. abs/2109.05433, 2021. [Online]. Available: https://arxiv.org/abs/2109.05433

[13] E. Commission, "Ai act," https://digital-strategy.ec.europa.eu/en/node/9745/printable/pdf, 2024, [Online; accessed May 22, 2024].

[14] F. Sovrano, S. Sapienza, M. Palmirani, and F. Vitali, "Metrics, explainability and the european ai act proposal," J, vol. 5, no. 1, pp. 126–138, 2022. [Online]. Available: https://www.mdpi.com/2571-8800/5/1/10

[15] A. Singhania, A. Unnam, and V. Aggarwal, "Grading video interviews with fairness considerations," CoRR, vol. abs/2007.05461, 2020. [Online]. Available: https://arxiv.org/abs/2007.05461

[16] C. Kim, J. Choi, J. Yoon, D. Yoo, and W. Lee, "Fairness-aware multimodal learning in automatic video interview assessment," IEEE Access, vol. 11, pp. 122 677–122 693, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3325891

[17] L. Chen, R. Zhao, C. W. Leong, B. Lehman, G. Feng, and M. E. Hoque, "Automated video interview judgment on a large-sized corpus collected online," in Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, October 23-26, 2017. IEEE Computer Society, 2017, pp. 504–509. [Online]. Available: https://doi.org/10.1109/ACII.2017.8273646

[18] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L. Morency, and M. E. Hoque, "Integrating multimodal information in large pretrained transformers," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 2359–2369. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.214

[19] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi, "Fairness without demographics through adversarially reweighted learning," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[22] I. Safran, G. Vardi, and J. D. Lee, "On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias," in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/d2dc4d6c7b102d05f111c02a32e7c6bc-Abstract-Conference.html

[23] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[24] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. J. Júnior, M. Madadi, S. Ayache, E. Viegas, F. Gürpinar, A. S. Wicaksana, C. C. S. Liem, M. A. J. van Gerven, and R. van Lier, "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," CoRR, vol. abs/1802.00745, 2018. [Online]. Available: http://arxiv.org/abs/1802.00745

[25] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America, 2005, pp. 1–16.

[26] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," IEEE Trans. Multim., vol. 16, no. 4, pp. 1018–1031, 2014. [Online]. Available: https://doi.org/10.1109/TMM.2014.2307169

[27] L. Chen, G. Feng, C. W. Leong, B. Lehman, M. P. Martin-Raugh, H. Kell, C. M. Lee, and S. Yoon, "Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016, Y. I. Nakano, E. André, T. Nishida, L. Morency, C. Busso, and C. Pelachaud, Eds. ACM, 2016, pp. 161–168. [Online]. Available: https://doi.org/10.1145/2993148.2993203

[28] J. Kim and W. Heo, "Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians," Inf. Technol. People, vol. 35, no. 3, pp. 861–878, 2022. [Online]. Available: https://doi.org/10.1108/ITP-04-2019-0173

[29] M. Langer, K. Baum, C. J. König, V. Hähne, D. Oster, and T. Speith, "Spare me the details: How the type of information about automated interviews influences applicant reactions," International Journal of Selection and Assessment, vol. 29, no. 2, pp. 154–169, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsa.12325

[30] W. Rahman, S. Mahbub, A. Salekin, M. K. Hasan, and E. Hoque, "Hirepreter: A framework for providing fine-grained interpretation for automated job interview analysis," in 2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021 - Workshops and Demos, Nara, Japan, September 28 - Oct. 1, 2021. IEEE, 2021, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ACIIW52867.2021.9666201

[31] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[32] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018. IEEE Computer Society, 2018, pp. 59–66. [Online]. Available: https://doi.org/10.1109/FG.2018.00019

[34] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 3319–3328. [Online]. Available: http://proceedings.mlr.press/v70/sundararajan17a.html

[35] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," CoRR, vol. abs/1908.05659, 2019. [Online]. Available: http://arxiv.org/abs/1908.05659

[36] R. Islam, H. Chen, and Y. Cai, "Fairness without demographics through shared latent space-based debiasing," in Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 12 717–12 725. [Online]. Available: https://doi.org/10.1609/aaai.v38i11.29167

[37] S. Lu, Y. Wang, and X. Wang, "Debiasing attention mechanism in transformer without demographics," in The Twelfth International Conference on Learning Representations, 2023.

[38] S. Zhao, D. Pascual, G. Brunner, and R. Wattenhofer, "Of non-linearity and commutativity in BERT," in International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021. IEEE, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN52387.2021.9533563

[39] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," Int. J. Uncertain. Fuzziness Knowl. Based Syst., vol. 6, no. 2, pp. 107–116, 1998. [Online]. Available: https://doi.org/10.1142/S0218488598000094

[40] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 2488–2498. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/905056c1ac1dad141560467e0a99e1cf-Abstract.html

[41] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali, "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey," Inf. Sci., vol. 615, pp. 238–292, 2022. [Online]. Available: https://doi.org/10.1016/j.ins.2022.10.013

[42] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," PloS one, vol. 10, no. 12, 2015.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[44] J. Louradour, "whisper-timestamped," https://github.com/linto-ai/whisper-timestamped, 2023.

[45] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: The dtw package," Journal of Statistical Software, vol. 31,

no. 7, p. 1–24, 2009. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v031i07

[46] L. Weytingh, L. Weytingh, J. Mohazzab, C. Wortmann, and B. B. Zaalberg, "Reimplementing the adversarially reweighted learning model by lahoti et al. (2020) to improve fairness without demographics," 2021. [Online]. Available: https://openreview.net/forum?id=P6-9f50PuMY

[47] J. Adler and I. Parmryd, "Quantifying colocalization by correlation: The pearson correlation coefficient is superior to the mander's overlap coefficient," Cytometry Part A, vol. 77A, no. 8, pp. 733–742, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.20896

[48] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, Eds. ACM, 2015, pp. 259–268. [Online]. Available: https://doi.org/10.1145/2783258.2783311

[49] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 3315–3323. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[50] R. Dwikifirdaus, "gender-classification," https://huggingface.co/rizvandwiki/gender-classification, accessed: 2024-05-26.

[51] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.

[52] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," CoRR, vol. abs/2009.07896, 2020. [Online]. Available: https://arxiv.org/abs/2009.07896

[53] D. Markova, L. Richer, M. Pangelinan, D. H. Schwartz, G. Leonard, M. Perron, G. Pike, S. Veillette, M. M. Chakravarty, Z. Pausova, and T. Paus, "Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence," Hormones and Behavior, vol. 81, pp. 84–96, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0018506X16301271

BIMASENA PUTRA received the BCS degree in computer science from Universitas Indonesia in 2024. He was a research student in the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan. His research interests include social signal processing, multimodal interaction, fairness, and machine learning.

KURNIAWATI AZIZAH (Member, IEEE) received the B.S. degree in informatics engineering from Bandung Institute of Technology (ITB), Bandung, Indonesia, in 1997, the M.Phil. degree in computer speech, text, and internet technology (CSTIT) from the University of Cambridge, Cambridge, U.K., in 2006, and the Ph.D. degree in computer science from Universitas Indonesia, Jakarta, Indonesia, in 2022.

She was an IT Consultant with MINCOM Indoservices, Jakarta, from 1998 to 2000, and with Switchlab, London, U.K., from 2000 to 2017. Since 2008, she has been a lecturer with the Faculty of Computer Science, Universitas Indonesia. Her research interests include deep learning, natural language processing (NLP), speech processing, and computer vision.

CANDY OLIVIA MAWALIM (Member, IEEE) received her B.S. in Computer Science from Institut Teknologi Bandung (ITB), Bandung, Indonesia. She received her M.S. and Ph.D. in the School of Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2019 and 2022, respectively. She was selected as a research fellow for young scientists DC1 (JSPS) in FY2020-2022. Since April 2022, she works as an assistant professor at the School of Information Science and Research Center for Biological Function and Sensory Information, JAIST. She is also on the education team of the ISCA special interest group of security and privacy in speech communication (SIG-SPSC) committee. Her main research interests are speech signal processing, hearing perception, voice privacy preservation, and machine learning.

IKHLASUL AKMAL HANIF is currently pursuing the BCS degree in the Faculty of Computer Science, Universitas Indonesia. He was a research student in the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan. His research interests include social signal processing, multimodal interaction, fairness, and machine learning.

SAKRIANI SAKTI (Member, IEEE) is the head of the Human-AI Interaction (HAI) Research Laboratory at the Nara Institute of Science and Technology (NAIST) in Japan, where she also serves as a full professor. Additionally, she is an adjunct professor at the Japan Advanced Institute of Science and Technology (JAIST) and the University of Indonesia, as well as a visiting research scientist at the RIKEN Center for Advanced Intelligent Project (RIKEN AIP) in Japan. In 2000, she was awarded the DAAD-Siemens Program Asia 21st Century Award to study Communication Technology at the University of Ulm, Germany, where she earned her MSc in 2002. From 2003 to 2011, she worked as a researcher at ATR and NICT in Japan. During this period, she pursued her Ph.D. with the Dialog Systems Group at the University of Ulm, completing it in 2008. She has participated in international collaborations, including the Asian Pacific Telecommunity Project (2003-2007) and speech-to-speech translation projects A-STAR and U-STAR (2006-2011). She was a Visiting Scientific Researcher at INRIA Paris-Rocquencourt, France, from 2015 to 2016. She currently serves as a committee member of the IEEE SLTC (2021-2026) and as an associate editor for IEEE/ACM TASLP (2020-2025), Frontiers in Language Sciences, and IEICE. Her research interests include deep learning, graphical model frameworks, spoken language processing and translation, and cognitive communication.

CHEE WEE (BEN) LEONG (Ph.D. Computer Science & Engineering) is a Director of AI Engineering at Educational Testing Service (ETS) in the U.S., where he directs, plan and implement prototyping of multimodal AI products and services in the educational domain and learning space. His research interests include multimodal modeling of affective states, evaluation and summarization of noncognitive skills, and multimodal data fusion for behavioral trait prediction.

SHOGO OKADA (Member, IEEE) directs the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan and is an associate professor at JAIST. He obtained his Ph.D. in 2008 from Tokyo Institute of Technology in Japan. In 2008 and 2011, he joined Kyoto University, Tokyo Institute of Technology, as an assistant professor. He joined IDIAP Research Institute in Switzerland as a visiting faculty member in 2014. His research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of the IEEE, ACM, JSAI, IEICE.

• • •