

Received 6 September 2023, accepted 2 October 2023, date of publication 19 October 2023, date of current version 7 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325891

RESEARCH ARTICLE

Fairness-Aware Multimodal Learning in Automatic Video Interview Assessment

CHANGWOO KIM^{1,*}, JINHO CHOI^{2,3}, JONGYEON YOON^{4,*}, DAEHUN YOO², AND WOJIN LEE⁵

¹Department of Artificial Intelligence, Dongguk University, Jung-gu, Seoul 04620, South Korea

²Genesis Laboratory, Jung-gu, Seoul 04538, South Korea

³Graduate School of AI, KAIST, Daejeon 34141, South Korea

⁴Department of Statistics, Dongguk University, Jung-gu, Seoul 04620, South Korea

⁵College of AI Convergence, Dongguk University, Jung-gu, Seoul 04620, South Korea

Corresponding author: Woojin Lee (wj926@dgu.ac.kr)

This work was supported in part by the Genesis Laboratory Company, and in part by the National Research Foundation of Korea (NRF) funded by the Korean Government through the Ministry of Science and Information and Communications Technology (MSIT) under Grant 2022R1F1A1074393.

*Changwoo Kim and Jongyeon Yoon contributed equally to this work.

ABSTRACT With the ever-growing reliance on Artificial Intelligence (AI) across diverse domains, there is an increasing concern surrounding the possibility of biases and unfairness inherent in AI systems. Fairness problems in automatic interview assessment systems, especially video-based automated interview assessments, have less been addressed despite their prevalence in the recruiting field. In this paper, we propose a method that resolves fairness problems in an automated interview assessment system that uses multimodal data as input. This is mainly done by minimizing the Wasserstein distance between two sensitive groups by introducing a regularization term. Subsequently, we employ a hyperparameter that can control the trade-off between fairness and accuracy. To test our method in various data settings, we suggest a preprocessing method that can manually adjust the underlying degree of unfairness in the training data. Experimental results show that our method presents state-of-the-art results in terms of fairness compared to previous methods.

INDEX TERMS Automatic interview assessment, multimodal, fairness, sensitive attribute, unfair assumption, Wasserstein distance, adversarial, representation learning.

I. INTRODUCTION

Video interviews were widely used even before, but their utilization has expanded significantly since the onset of the COVID-19 pandemic to reduce physical interaction. Among various methods, asynchronous video interviews (AVIs) are currently the most prevalent method [1]. In AVIs, the applicants pre-record their responses to predetermined questions, then they are subsequently assessed by an interviewer who evaluates the recorded videos. AVIs allow companies to interview more applicants while providing applicants with the flexibility to participate in interviews at their preferred time and location. Because of these conveniences, AVIs are now widely used online interview methods even after the pandemic. Although AVIs can provide opportunities to a larger

pool of applicants, the task of evaluating a substantial number of recorded videos with a limited number of interviewers can be inefficient and time-consuming. To address the inherent challenges associated with interview assessments relying solely on human, automated video interview assessment systems that combine powerful machine learning algorithms have recently emerged [2], [3], [4], [5]. An automated video interview assessment system gathers human-labeled interview scores by reviewing pre-recorded interview videos of the applicants. Subsequently, the machine learning model is trained using pre-recorded interview videos as inputs, aiming to predict the appropriate interview scores for each individual.

As such, machine learning is widely applied in today's recruitment process, however, there is a growing concern about whether the interview results predicted by such models are fair. In machine learning, fairness refers to making

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

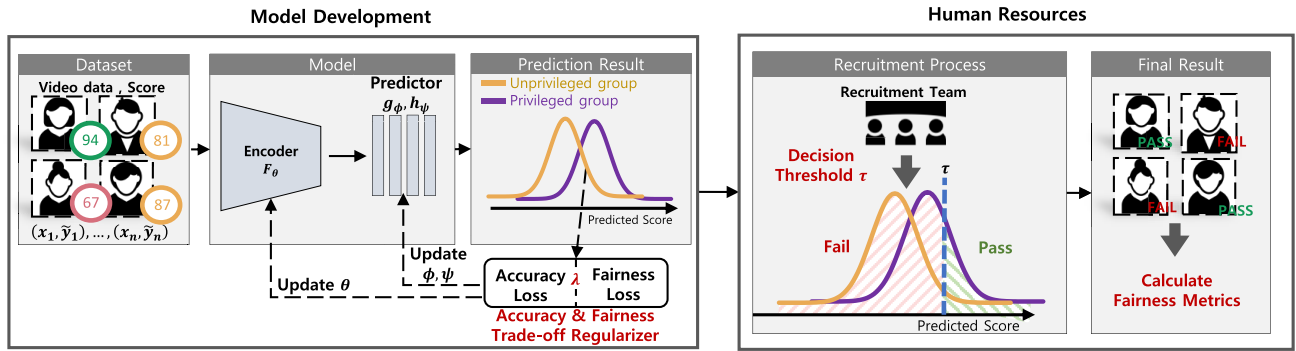


FIGURE 1. Problem setting of our paper. Our problem setting can be divided into two sections: the model development and the human resources section. In the model development section, the dataset consists of video data x and a continuous variable score y . Giving x as an input to encoder F_θ , predictor g_ϕ predicts the score using the encoded input. In the human resources section, the recruitment team decides the decision threshold τ , to decide whether the candidates pass or fail the interview, based on their predicted score.

non-discriminatory decisions against certain groups or sensitive attributes such as gender, age, or race. Concerningly, it has been observed that machine learning-based recruitment models that do not consider fairness tend to make decisions based on sensitive attributes. In other words, they advantage privileged group or disadvantage unprivileged group based on unwanted bias. Note that the term ‘privileged group’ refers to a group that receives advantages due to unwanted bias, while the term ‘unprivileged group’ refers to a group that experiences disadvantages due to unwanted bias. One notable example is Amazon’s artificial intelligence (AI) hiring system, which showed a tendency to discriminate against women in software development and technical positions [6]. In terms of legal aspects, the Equal Employment Opportunity Commission in the United States prohibits discrimination based on factors such as gender, race, and skin color during hiring. In this social circumstance, research has been conducted on fairness in automated recruitment processes, such as resume screening [7], [8]. However, the issues of fairness have been less explored in the context of automated video interview assessment systems, despite their increasing prevalence in real-world applications.

The difficulty of addressing fairness issues in video-based interview processes stems from the following reasons. Firstly, fairness problems in machine learning usually deal with binary labels (e.g., pass or fail), aiming to prevent discrimination between privileged and unprivileged groups in decision-making. However, in the case of automated video interview assessment problems, an automatic model aims to predict a continuous variable score, rather than performing binary classification. As shown in the left side of Figure 1, candidates’ label score \tilde{y} exists in continuous form. Hence, an automatic model predicts continuous interview scores for each candidate, which are later used by the recruitment team. The recruitment team assigns a threshold to the continuous interview scores assigned by the automatic model, thereby making the label in binary form. Therefore, this can be seen as a combined problem of regression and classification. This makes it challenging to directly apply existing fairness machine learning algorithms that are based

on binary classification. To the best of our knowledge, there has been no fairness research that considered the conversion from numeric variable to binary variable.

Secondly, majority of fairness research is based on optimization algorithms such as linear programming or convex optimization [9], [10], [11] and used tabular dataset which has low data complexity. This makes it difficult to apply them to large-scale multimodal datasets such as videos due to computational costs. Also, the data complexity is not easily handled by those methods.

Finally, there exists a problem due to the trade-off between a model’s accuracy and fairness. When models are trained to improve fairness, they often exhibit a decrease in accuracy and vice versa. A major reason for this trade-off is some models may exhibit high accuracy for specific groups, but accuracy could be low for other groups. To enhance fairness, efforts should be made to reduce performance disparities among groups, but this could potentially lead to an overall performance degradation. From a practical application perspective, it is important to have high accuracy while avoiding fairness-related issues in any situation. Therefore, there is a demand for models that can adjust fairness and accuracy according to various situations.

To address fairness in the automated video interview assessment system, we present a method that effectively handles these problems. First, we propose a loss function that can maintain fairness for arbitrary thresholds used to determine the interview outcomes. In this point of view, we provide a theoretical explanation that reducing the Wasserstein distance [12], [13] between output distributions can be an effective method to improve fairness in our problem settings. Wasserstein distance is widely utilized in the field of deep learning for various applications. Wasserstein distance’s benefits include robustness, intuitive interpretation, and stable learning in generative models. It supports optimal matching, aids in distribution interpolation, and finds use in diverse fields such as generative modeling (WGAN [14]) and data analysis due to its ability to accurately measure distribution differences. Additionally, we have developed a deep learning-based architecture and optimization to

effectively apply fairness algorithms to multi-modal data, aiming to reduce the aforementioned Wasserstein distance and learn a fair representation by using an adversarial training approach. We also employ a hyperparameter to address the trade-off between fairness and accuracy. In Figure 1 this hyper-parameter is denoted by λ . By tuning λ , the user of this model can control the importance of fairness loss in the total loss thereby controlling the trade-off between accuracy and fairness. This control allows for the customization of the balance between accuracy and fairness, which can be adjusted according to various situations.

To evaluate the performance of the proposed method in different fairness scenarios, comparative experiments with existing fairness methods are conducted. For example, datasets with various degrees of fairness, which were created by intentionally manipulating the existing dataset, were used to test our method. The number of privileged group and unprivileged group in target variable were manipulated to control the degrees of unfairness. It is known that the fairness of a machine learning algorithm is highly dependent on the degree of unfairness present in the training data. The more unfair the training data is, the more likely the model is to make discriminatory predictions. To test the proposed method on various fairness settings, we suggest a preprocessing method that incorporates a sampling method, taking into account both the label and the sensitive attribute. This method allows for manual adjustment of the underlying degree of unfairness in training data.

Experiments were conducted using two different datasets in various fairness scenarios created by the suggested preprocessing method. Our approach has demonstrated superior performance in terms of fairness while minimizing the decrease in accuracy compared to existing methods across various unfair scenarios.

In summary, the contributions of our work are as follows:

- We conduct fairness experiments using real-world video interview assessment data.
- We propose a novel algorithm that effectively addresses fairness issues in the automated video interview assessment system.
- We propose an evaluation process to assess whether the fairness of the model can be maintained even when utilizing an unfair training dataset.
- Our algorithm shows the best trade-off curve, minimizing performance drop across various situations when compared to existing methods.

II. BACKGROUND AND RELATED WORK

A. AUTOMATED ASSESSMENT OF JOB INTERVIEW AND FAIRNESS

The goal of automatic video interview assessment is to predict the hiring recommendation score based on verbal (e.g., speech content) and non-verbal (e.g., loudness, tone of voice, body gestures, eye gaze, and facial expression) behaviors in job interview videos, without any human intervention. Due to the advancements in deep learning, automated

video interview systems have been actively researched and developed. Nguyen et al. [15] proposed a method that predicts the job interview hirability score solely on the non-verbal behaviors of both the candidates and the interviewer. Naim et al. [16] use both verbal and non-verbal features to assess soft skills such as friendliness, engagement, and hiring recommendations. In addition, a system that recognizes the non-verbal behaviors of candidates was proposed [2] and platforms that evaluate candidates using various features were suggested [3], [4].

As such, an automated interview assessment system is being extensively researched. However, the employment of AI technologies in the recruitment field has raised substantial concerns about the fairness of their results. Hunkenschroer et al. [17] conducted an ethical examination of AI-based recruitment, focusing on human rights perspectives. The authors contend that the ethical consequences of AI-based recruitment heavily depend on the particular contexts in which these tools are utilized.

To address fairness in the field of recruitment, much research was conducted. Pessach et al. [7] introduced an approach to develop a fair AI algorithm in an unbalanced data setting, in which the sample size of unprivileged groups is considerably smaller compared to that of privileged groups. Additionally, Pena et al. [8] have proposed a method to make fair decisions in an automated interview assessment system using multimodal AI.

The aforementioned methods are conducted using a dataset comprising simple information such as the candidate's personal detail (e.g., age, residence). However, there has been insufficient research on the fairness problem yet on the widely used video-based interview assessment method. The difficulty of addressing fairness in this method can be summarized in three factors. First, fairness problems in machine learning typically pertain to binary labels, while interview assessment problems involve continuous predictions from models, which makes it challenging to directly apply existing methods. Second, most fairness research is based on optimization algorithms, which are computationally expensive and hinder their application in large-scale multimodal scenarios. Finally, the trade-off between model accuracy and fairness is another obstacle in addressing fairness with the video-based assessment method. It is well-known that when models are trained with an emphasis on fairness, they often experience a decrease in accuracy. Excessive consideration of fairness can significantly decrease accuracy, leading to difficulties in addressing fairness in video-based assessment.

Considering the fact that the video-based assessment method is widely used in the real world, it is important to improve fairness in this field.

B. FAIRNESS DEFINITIONS

In fairness problems, it is commonly stated that a model is considered fair if its decision does not depend on sensitive attributes such as gender, age, or race. Attributes in the dataset

can be divided into three parts X , S , and Y . X is a collection of non-sensitive attributes such as input and S denotes a sensitive attribute. In this work, we set S as binary attribute, $S \in \{0, 1\}$, where $S = 1$ means *privileged* and $S = 0$ means *unprivileged*. Y is a label that we are trying to predict. Y is often set as binary attribute, $Y \in \{0, 1\}$, where $Y = 1$ means positive outcome like *pass* and $Y = 0$ means negative outcome like *fail*. So we call that decision fair if predicted label \hat{Y} and S are independent.

To quantify the discrimination between groups, various fairness metrics have been proposed to measure the fairness of a machine learning model.

Demographic Parity (DP) is a metric that is utilized widely to assess fairness, specifically focusing on the disparity in acceptance rates between different groups based on sensitive attributes. It enforces statistical independence between the \hat{Y} and S . The classification model is considered fair under the DP criterion if DP is close to zero. DP is formulated as follows:

$$DP = |Pr_X(\hat{Y} = 1|S = 0) - Pr_X(\hat{Y} = 1|S = 1)|. \quad (1)$$

However, using DP as a measure to guarantee fairness has potential drawbacks. One limitation of this metric is that an entirely accurate classifier may be perceived as unfair when the proportions of positive samples ($Pr_X(Y = 1)$) differ among groups. Moreover, in pursuit of DP, two similar individuals might receive unequal treatment purely based on their affiliation with different groups.

To address this problem, the *Equal Opportunity (EO)* criterion incorporates the true label information (Y) in addition to the predicted label \hat{Y} . In other words, EO assesses whether the model provides an equal chance of correctly predicting the positive instances for all groups. The classification model is considered fair under the EO criterion if EO is close to 0. The equation is as follows:

$$EO = Pr_X(\hat{Y} = 1|S = 0, Y = 1) - Pr_X(\hat{Y} = 1|S = 1, Y = 1) \quad (2)$$

C. MECHANISMS FOR IMPROVING FAIRNESS

Mechanisms for improving fairness can be divided into three categories: pre-processing, in-processing, and post-processing.

Pre-processing mechanisms recognize that biased, discriminatory, or imbalanced distributions of the sensitive attribute are the cause of fairness issues. Therefore, pre-processing mechanisms involve modifying the training data before using it in algorithms [18], [19], [20], [21], [22], [23]. The most commonly used method is to either flip or modify the dependent variable or otherwise change the distribution of independent variables. This involves changing the labels of some instances or modifying feature representations to ensure that the classifier is fair [18], [19], [20].

In-processing mechanisms aim to modify the training algorithms to incorporate fairness considerations during the

training phase. Krasanakis et al. [24] learn the weight of samples in a way that reduces biases. Yan et al. [25] aims to achieve an equitable distribution of the population across various sensitive groups. Zhang et al. [26] used adversarial learning to penalize the model if the sensitive variable is predictable from the dependent variable. The most widely used method is adding fairness-related penalty terms in the objective function [9], [27], [28], [29]. For example, adding mutual information regularization terms between a sensitive attribute S and predicted outcome \hat{Y} into the objective function will guide the model to make predictions independent of the sensitive attributes.

Post-processing mechanisms seek to improve fairness by modifying the output scores of the classifier. For example, it computes the threshold value where the privileged group and unprivileged group are both fairly classified [30], [31], [32].

Three different types of mechanisms for improving fairness have pros and cons. Pre-processing mechanisms can be used in any classification algorithm since they are applied to the original dataset, not the model itself. However, they have high uncertainty in the model accuracy since they are not tailored for specific classification algorithms. Like pre-processing mechanisms, post-processing mechanisms have an advantage in their applicability to any classification algorithms. However, as they are utilized in the latter stages of the training process, their outcomes might be less optimal compared to those of other types of mechanisms. In-processing mechanisms have an advantage in that they have a high probability of obtaining superior results with respect to fairness and accuracy. Nonetheless, these mechanisms have not been widely researched on high-dimensional data, since previous works were focused on numerical datasets.

In this paper, we utilize the in-processing method among three approaches to construct a fair model for automatic video interview assessment. We built a fairness model for an automated video interview assessment system by employing a deep learning-based approach, which can incorporate the multimodal characteristics of the data.

III. PROBLEM SETTINGS AND THEORETICAL BACKGROUND

This paper aims to address fairness issues in automatic interview processes. As explained in the previous section, we define the input space as $\mathcal{X} \in \mathbb{R}^d$, where $X \in \mathcal{X}$ represents the input data and $S \in \{0, 1\}$ denotes the sensitive attribute, and $Y \in \{0, 1\}$ represents the label. We use a binary classifier $g : \mathcal{X} \rightarrow \{0, 1\}$ to determine whether a candidate *passes* or *fails* the interview.

However, in the context of automatic interview assessment, we assume that we have access to input data X , but lack information about the binary label Y of the candidate. Instead, we only have access to their continuous interview scores, denoted as $\tilde{Y} \in [0, 1]$. Our objective is to build an interview assessment model that predicts the interview scores for each candidate using X and \tilde{Y} . Subsequently, the predicted label \hat{Y}

is determined based on the distribution of the predicted score values from candidates.

Therefore, in this paper, we assume that the binary classification prediction is based on the score function $\eta : \mathcal{X} \rightarrow [0, 1]$, which estimates the predicted score for each sample $x_i \in X$. The classification decision is made based on a threshold τ that converts the output of the score function η into a binary output. We consider a group of threshold classifiers $\{g_\tau\}_{\tau \in (0,1)}$ that uses the classification rule

$$g_\tau = 1_{\eta(x) > \tau}(x). \quad (3)$$

The threshold classifier g_τ predicts *pass* if and only if the output of the score function η is bigger than τ . The fairness metric DP gap (ΔDP) can be defined given g_τ as follows:

$$\Delta DP(g_\tau) = |Pr_X(\hat{Y} = 1 | S = 1) - Pr_X(\hat{Y} = 1 | S = 0)|. \quad (4)$$

In automated interviews, since the threshold for acceptance or rejection will be determined at a later stage, we aim to utilize a metric in this paper that ensures fairness for all thresholds. This metric is called *Strong Pairwise Demographic Disparity (SPDD)* and in this setting, the SPDD(η) can be defined as follows:

$$SPDD(\eta) = \mathbb{E}_{\tau \sim U((0,1))} \Delta DP(g_\tau), \quad (5)$$

where U denotes uniform distribution and η is a score function. In this paper, we plan to utilize SPDD as a basic metric to measure fairness, in the context of automated interviews.

A. THE WASSERSTEIN DISTANCE

The Wasserstein distance, also known as Earth Mover's Distance (EMD), is a measure of the distance between two probability distributions. It is calculated by finding the minimum cost required to transform one distribution into another. The Wasserstein distance is currently being researched from a fairness perspective as it can measure the differences in predictions between different groups in group fairness problems [33], [34].

Given two probability density functions (PDF) p_0 and p_1 both on latent space \mathcal{Z} , satisfying $\int c(x, y) dp_i < \infty$ for $i = 0, 1$, the optimal transport map T is the one that minimizes the total transportation cost

$$T^* = \arg \min_T \int_{\mathcal{Z}} c(z, T(z)) dp_0(z), \quad (6)$$

under the condition $T_{\#} p_0 = p_1$, meaning that T push forwards p_0 to p_1 , where \mathcal{T} is the set of transportation maps, and $c : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a cost function.

To find the optimal transport map more efficiently, we can reformulate the optimal transport problem in the search for an optimal joint PDF

$$\min \int_{\mathcal{Z} \times \mathcal{Z}} c(p_0, p_1) d\gamma(p_0, p_1), \quad (7)$$

over $\gamma \in \Pi(p_0, p_1) = \{\gamma | \pi_{\#} \gamma = p_i, i = 0, 1\}$, where π_i denote the marginal projections $\mathcal{Z} \times \mathcal{Z} \xrightarrow{Z} \mathcal{Z}$. In this setting, we can define the k -Wasserstein distance as

$$W_p(p_0, p_1) = \inf_{\gamma \in \Pi(p_0, p_1)} \left(\int_{\mathcal{Z} \times \mathcal{Z}} d(p_0, p_1)^k d\gamma(p_0, p_1) \right)^{\frac{1}{k}}, \quad (8)$$

where $k \geq 1$.

Wasserstein distance is computationally expensive in many cases, however, for the 1-Wasserstein distance, there is a simple approach that can calculate the distance [12]

$$W_1(p_0, p_1) = \sup \left(\int_{\mathcal{Z}} f dp_0 - \int_{\mathcal{Z}} f dp_1 : \|f\|_L \leq 1 \right), \quad (9)$$

where the condition $\|f\|_L \leq 1$ requires that f is 1-Lipschitz as a function from \mathcal{Z} to \mathbb{R} with respect to the metric d .

Wasserstein distance has gained popularity in deep learning as a key component in loss functions. It offers advantages over other discrepancy measures, including total variation distance, Kullback-Leibler divergence, and Jensen-Shannon divergence. Unlike the aforementioned dissimilarity measures, Wasserstein distance considers the underlying geometry and assigns a finite distance value even when two distributions do not have overlapping support [13], [35].

B. FAIRNESS AND THE WASSERSTEIN DISTANCE

We now explain the relationship between fairness criteria SPDD, the most suitable metric for automated interview scenarios, and Wasserstein distance.

Let S_1 and S_2 be two random variables that take values in the interval $\Omega = [0, 1]$. For $i = 0$ and 1 , let F_i represent the cumulative distribution function (CDF) of S_i . If we set μ_i be the distribution on Ω induced by the variable S_i , then we have

$$W_1(\mu_0, \mu_1) = \int_{\Omega} |F_0(\tau) - F_1(\tau)| d\tau. \quad (10)$$

Importantly, we can discover a direct relationship between the Wasserstein distance in one-dimensional distributions and the fairness concepts of SPDD. Let's examine the conditional distributions \mathcal{D}_i that are dependent on sensitive groups, represented as $\mathcal{L}(\mathcal{X} | S = i)$, where i takes values 0 and 1. Assuming we have a trained score function $\eta : \mathcal{X} \rightarrow [0, 1]$, we can represent distributions as the push-forward distributions

$$\mu_i = \eta_{\#} \mathcal{D}_i, i = 0, 1. \quad (11)$$

Then we can obtain

$$\begin{aligned} W_1(\mu_0, \mu_1) &= \int_0^1 \left| \Pr_{X \sim \mathcal{D}_0} (\eta(X) \leq \tau) - \Pr_{X \sim \mathcal{D}_1} (\eta(X) \leq \tau) \right| d\tau \\ &= \int_0^1 \left| \Pr_{X \sim \mathcal{D}_0} (\eta(X) > \tau) - \Pr_{X \sim \mathcal{D}_1} (\eta(X) > \tau) \right| d\tau \\ &= \int_0^1 \left| \Pr_X(\hat{Y}_\tau = 1 | S = 0) - \Pr_X(\hat{Y}_\tau = 1 | S = 1) \right| d\tau \\ &= \mathbb{E}_{\tau \sim U((0,1))} \Delta DP(h_\tau), \end{aligned}$$

and the last term is precisely the SPDD that we defined previously.

Based on above mentioned observations, we have noticed that controlling the 1-Wasserstein between distributions of score values effectively regulates the SPDD. Now, utilizing this property, we propose approaches in the context of automated interview problems that can reduce the Wasserstein distance in output distributions between privileged and unprivileged groups.

IV. PROPOSED METHODS

A. NOTATIONS

We consider classification tasks in which $\mathcal{X} \in \mathbb{R}^d$ is an input space and $\mathcal{Z} \in \mathbb{R}^k$ is a latent space. The encoder $F_\theta : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by θ , yields the representation from input. Z can be expressed as $Z = F(X)$. Note that since we are dealing with three types of data (video (V), text (T), and audio (A)), we have inputs, encoders, and representations for each type of data. Therefore, X , F , and Z each consists of (X_V, X_T, X_A) , (f_V, f_T, f_A) , and (Z_V, Z_T, Z_A) . Then, we concatenate representations of each input. As a result, concatenated representation Z_C can be expressed as

$$\begin{aligned} Z_C &= [Z_V, Z_T, Z_A] \\ &= [f_V(X_V), f_T(X_T), f_A(X_A)]. \end{aligned} \quad (12)$$

Using concatenated representation Z_C , regressor $g_\phi : \mathcal{Z} \rightarrow [0, 1]$ and adversary $h_\psi : \mathcal{Z} \rightarrow \{0, 1\}$, each parameterized by ϕ and ψ , predicts \tilde{Y} and S respectively.

B. IMPROVING FAIRNESS VIA ADVERSARIAL TRAINING

In this paper, we try to use the in-processing approach for fairness, as we aim to encode representations that are independent of sensitive attributes. We do this by training a classifier called adversary h , that predicts sensitive attribute S using encoded representation Z , encoded by encoder F , as input. Note that Z is equivalent to $F(X)$. Since S is binary we can define

$$\begin{aligned} D_{\theta, \psi}(X, S) &= \mathbb{E}_{X, S} [S \cdot \log(h(F(X))) \\ &\quad + (1 - S) \cdot \log(1 - h(F(X)))] \end{aligned} \quad (13)$$

that is the negative of the binary cross-entropy loss. The adversary's parameters ψ are parameterized to maximize D , while the encoder's parameters θ are parameterized to minimize D [36]. The minimax problem can be shown as follows:

$$\min_{\theta} \max_{\psi} D_{\theta, \psi}(X, S). \quad (14)$$

C. IMPROVING FAIRNESS VIA WASSERSTEIN DISTANCE

The purpose of this work is to train a model with a regressor that is fair with respect to sensitive attributes. As mentioned earlier, the reduction of Wasserstein distance is the key to achieving fairness in our settings. Therefore, we utilize the 1-Wasserstein distance between sensitive groups as a

regularization term in our objective function (loss function) of the regressor.

However, there is a problem that the exact computation of true label distributions is intractable. For this reason we use empirical distributions $\hat{\mu}$, defined by

$$\hat{\mu}_a = \frac{1}{|B_s|} \sum_{i \in B_s} \delta_{g_\phi \circ f_\theta(X_i)} \quad (15)$$

for $s = 0, 1$, where B_s is a subset of index set

$$I_s = \{i = 1, \dots, n : s_i = s\} \quad (16)$$

and δ_p is the Dirac measure centered at $p \in \mathbb{R}$. Note that each δ follows delta distributions, thereby the distributions in (15) are uniform mixtures of delta distributions centered at the sets B_s of samples drawn from the respective underlying distributions ($s = 0, 1$). Using this empirical distribution, we define the regularization term

$$\mathcal{L}_W = W_1(\hat{\mu}_0, \hat{\mu}_1). \quad (17)$$

Note that as mentioned in III, minimization of (17) can be viewed as stochastic minimization of SPDD.

Let, $\hat{\mu}_0 = \frac{1}{m} \sum_{i=1}^m \delta_{p_i}$ and $\hat{\mu}_1 = \frac{1}{m} \sum_{i=1}^m \delta_{q_i}$, where $p_i, q_i \in \mathbb{R}$ for $i = 1, \dots, m$. Let $\rho : \mathbb{R}^m \rightarrow \mathbb{R}^m$ be a sorting function such that given a vector $\mathbf{r} = (r_1, \dots, r_m)$, outputs $\rho(\mathbf{r}) = (r_{\sigma(1)}, \dots, r_{\sigma(m)})$ where $\sigma : 1, \dots, m \rightarrow 1, \dots, m$ is a rearrangement of indices such that $1 \leq i < j \leq m$ implies $r_{\sigma(i)} \leq r_{\sigma(j)}$. In this setting, the optimal coupling γ^* is simply the assignment $\rho(\mathbf{p})_i \mapsto \rho(\mathbf{q})_i$ for each $i = 1, \dots, m$. Therefore, the 1-Wasserstein distance between $\hat{\mu}_0$ and $\hat{\mu}_1$ is given by

$$W_1(\hat{\mu}_0, \hat{\mu}_1) = \frac{1}{m} \sum_{i=1}^m |\rho(\mathbf{p})_i - \rho(\mathbf{q})_i|, \quad (18)$$

where m is the size of the mini-batch. Using equation (18), we minimize the 1-Wasserstein distance between two one-dimensional distributions, in our setting distance between the privileged group and the unprivileged group.

Even though it is a simple method, estimating empirical distributions like (18) is useful in computing empirical versions of the 1-Wasserstein distance. This is because W_1 distance can be accurately computed using a simple closed-form expression for one-dimensional empirical distributions with an equal number of point masses [37].

It is important that the number of point masses has to be the same for the distributions. In our implementation, this condition will always be satisfied, since B_0 and B_1 are batches for stochastic gradient descent, and their size is fixed to constant m .

D. OBJECTIVE FUNCTION

The adversary h_ψ and regressor g_ϕ seeks to minimize its loss on predicting S and \tilde{Y} respectively from Z . Note that \tilde{Y} is a continuous interview scores of candidates that we have access to. Also encoder f_θ seeks to encode inputs to minimize the regressor's loss and maximize the adversary's loss. Let L_{Reg}

Algorithm 1 Training Procedure for the Proposed Method

Input: Labeled dataset $(X_V, X_A, X_T) \in X, S$ and \tilde{Y}
Output: Trained encoder F_θ , trained regressor g_ϕ and adversary h_ψ

- 1: **Initialize the multimodal model with random parameters.**
- 2: Init encoder $(f_V, f_A, f_T) \in F_\theta$
- 3: Init regressor g_ϕ that predicts *score* \tilde{Y}
- 4: Init adversary h_ψ that predicts sensitive attribute S
- 5: **Set hyperparameters**
- 6: WD loss weight λ_W , Adversarial weight λ_{Adv}
- 7: Batch size m , epochs e , Learning rate for regressor η_{Reg} and adversary η_{Adv}
- 8: **for** each epoch **in** e **do**
- 9: **for** each batch B , where $|B_{S=0}| + |B_{S=1}| = m$ **do**
- 10: Calculate WD loss about sensitive attribute:
 - 11: Extracts representations in latent space $F_\theta(B_{S=0})$ and $F_\theta(B_{S=1})$ from each batch;
 - 12: Compute regressor output $g_\phi(F_\theta(B_{S=0}))$ and $g_\phi(F_\theta(B_{S=1}))$ using representations as inputs;
 - 13: Sort $g_\phi(F_\theta(B_0))$ with ρ such that $\rho(g_\phi(F_\theta(X)))_i \leq \rho(g_\phi(F_\theta(X)))_{i+1}, \forall X \in B_{S=0}, 1 \leq i < m$;
 - 14: Sort $g_\phi(F_\theta(B_1))$ with ρ such that $\rho(g_\phi(F_\theta(X)))_i \leq \rho(g_\phi(F_\theta(X)))_{i+1}, \forall X \in B_{S=1}, 1 \leq i < m$;
 - 15: $k = \min(\text{len}(B_{S=0}), \text{len}(B_{S=1}))$
 - 16: $L_W = \frac{1}{k} \sum_{i=1}^k |\rho(g_\phi(F_\theta(B_0)))_i - \rho(g_\phi(F_\theta(B_1)))_i|$
- 17: Calculate Supervised loss:
 - 18: $L_{MSE} = \sum_{(X_i, \tilde{Y}_i) \in B} l(g_\phi(F_\theta(X_i), \tilde{Y}_i))$
- 19: Calculate total loss:
 - 20: $L_{Reg} = L_{MSE} + \lambda_W L_W$
- 21: Update ϕ and θ with gradient descent;
 - 22: $\phi \leftarrow -\eta_{Reg} \nabla_\phi \mathcal{L}_{Reg}$
 - 23: $\theta \leftarrow -\eta_{Reg} \nabla_\theta \mathcal{L}_{Reg}$
- 24: Calculate adversary loss:
 - 25: Extracts representations in latent space $F_\theta(B)$ from each batch;
 - 26: Compute adversary output $h_\psi(F_\theta(B))$ using representations as inputs;
 - 27: $L_{Adv} = \lambda_{Adv} \sum_{(X_i, S_i) \in B} l(h_\psi(F_\theta(X_i), S_i))$
- 28: Update ψ with gradient descent:
 - 29: $\psi \leftarrow -\eta_{Adv} \nabla_\psi L_{Adv}$
- 30: Update θ with reverse gradient descent:
 - 31: $\theta \leftarrow -(-\eta_{Adv} \nabla_\theta L_{Adv})$
- 32: **end for**
- 33: **end for**

denote a suitable regression loss and L_{Adv} denote a suitable classification loss. For L_{Reg} , we use mean squared error with (17) added as a regularization term. So L_{Reg} can be defined as

$$L_{Reg}(X, \tilde{Y}) = \mathbb{E}_{X, \tilde{Y}} \left\| g(F(X)) - \tilde{Y} \right\|_2^2 + \lambda_W \cdot \mathcal{L}_W. \quad (19)$$

Note that λ_W is a hyper-parameter for balancing fairness and accuracy.

For L_{Adv} , we use the negative of the binary cross-entropy loss (13), defined in Section IV-B.

$$L_{Adv}(X, S) = D_{\theta, \psi}(X, S) \quad (20)$$

We can train the model by optimizing the following two equations sequentially.

$$\min_{\phi, \psi} L_{Reg}(X, \tilde{Y}) + L_{Adv}(X, S) \quad (21)$$

TABLE 1. Underlying bias of each dataset. We measured the inherent bias in the dataset by measuring DP_{data} and $SPDD_{data}$. It can be observed that the degree of unfairness increases as the value of α increases.

Train dataset	HR dataset		FI dataset	
	DP_{data}	$SPDD_{data}$	DP_{data}	$SPDD_{data}$
original	0.000	0.011	0.000	0.006
$\alpha = 2$	0.167	0.084	0.167	0.044
$\alpha = 3$	0.253	0.121	0.250	0.064
$\alpha = 4$	0.300	0.142	0.300	0.076

$$\min_{\theta} L_{Reg}(X, \tilde{Y}) - L_{Adv}(X, S) \quad (22)$$

Algorithm 1 outlines the detailed training procedure for the proposed method.

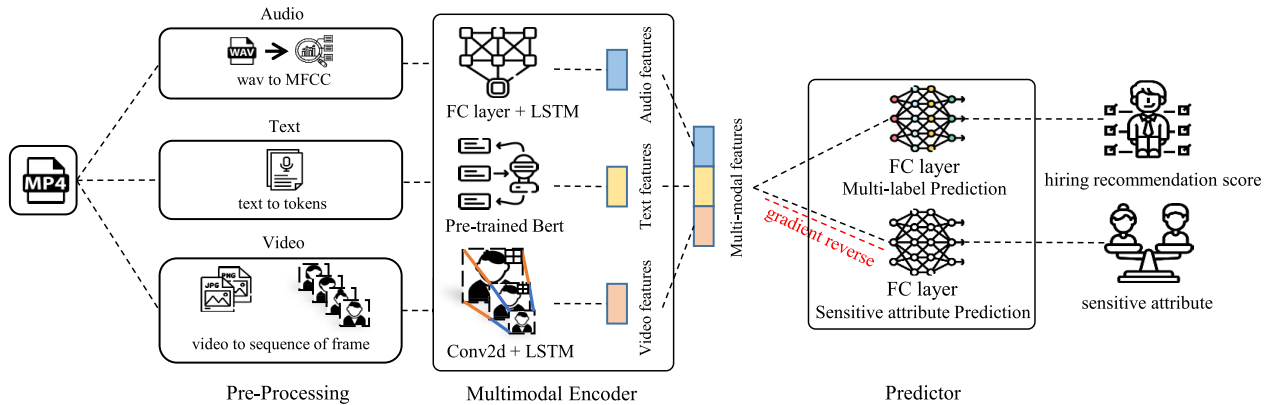


FIGURE 2. Overall architecture of the proposed model. The model aims to make fair predictions using multiple modalities, including video, audio, and text input. Each modality pre-processing step extracts relevant features. The preprocessed features are used as inputs for each model in the multimodal architecture. The encoded features obtained through the model are used to predict scores using a multi-label predictor, and gender is predicted using a sensitive attribute classifier.

V. EXPERIMENTAL SETTINGS

A. DATASET

1) HIRING RECOMMENDATION DATASET

The Hiring Recommendation (HR) dataset consists of real-world job interview videos with candidates, speaking in Korean language. In each video, participants respond to a pre-defined question within 90 seconds. These videos are annotated with the gender of the participant and the *hiring recommendation score* [16] assigned to them as a measure of their likelihood of securing a job offer. The target label of *hiring recommendation score* is provided by three interview experts with a minimum of 20 years of experience. To ensure the validity of the label, the same videos are assigned to annotators, and the average *hiring recommendation score* is used. The resulting dataset contains data from more than three thousand interviewees. We employed a five-fold cross-validation approach for training and evaluation, where 80% of the total data is utilized for training and the remaining 20% for the test dataset. In the HR dataset, target label \tilde{Y} represents the *hiring recommendation score*, and sensitive attribute S represents *gender*. Input variable X is composed of audio, video, and text data from the candidate’s interview video.

2) FIRST IMPRESSIONS DATASET

To ensure that our fairness approach performs well with another dataset, we additionally used the First Impressions (FI) dataset [38]. This dataset was also used in the 2017 Chalearn Lab challenge at Computer Vision and Pattern Recognition (CVPR). It consists of 10,000 video clips extracted from over 3,000 different high-definition (HD) YouTube videos of people speaking in English while facing a camera. These videos are labeled with *gender* and continuous variable *interview* that indicate whether the subject should be invited for a job interview. The 10,000 clips are split into a 60% training set, a 20% validation set, and a 20% testing set. Along with the video clips, we also used their transcriptions provided by Chalearn Lab in our study. In the FI dataset, the

target variable \tilde{Y} is *Interview* and the sensitive attribute S is *gender*. Input variable X is composed of audio, video, and text data from the YouTube video.

3) TARGET SETTINGS

The target variable \tilde{Y} is termed *score* in both datasets. Data samples with a *score* of 0.5 or higher, are categorized as the high-scoring candidates ($score \geq 0.5$), while data samples with a score below 0.5, are the low-scoring candidates ($score < 0.5$). For the sensitive attribute S , female is set as a *privileged (priv.)* group ($S = 1$) and male as *unprivileged (unpriv.)* group ($S = 0$).

B. DEGREE OF UNFAIRNESS AND DATASET MANIPULATION

In an automated job interview assessment, the fairness of the model is significantly influenced by the proportion of different demographic group members among high-scoring and low-scoring candidates within the training data. Therefore, we present α , an indicator that can quantitatively measure the difference in the proportion of each sensitive attribute between the high-scoring and low-scoring groups, as follows:

$$\alpha_{high} = \frac{Pr(S = 1 | \tilde{Y} \geq \tau)}{Pr(S = 0 | \tilde{Y} \geq \tau)} \quad (23)$$

$$\alpha_{low} = \frac{Pr(S = 0 | \tilde{Y} < \tau)}{Pr(S = 1 | \tilde{Y} < \tau)}$$

$$\alpha = \frac{\alpha_{high} + \alpha_{low}}{2},$$

where α_{high} is the proportion of the number of *priv.* to the number of *unpriv.* in the high-scoring group, and α_{low} represents the proportion of the number of *unpriv.* to the number of *priv.* in the low scoring group. The α is the average value of α_{high} and α_{low} . A higher α value indicates significantly more candidates with a specific sensitive attribute in the high and low-scoring groups. Tables 2 and 3 illustrate the difference in proportions of candidates with

TABLE 2. Distribution of target(score) and sensitive attributes(gender) in the HR unfair dataset (threshold = 0.5). Samples with a score of 0.5 and higher are classified as a high-scoring group, and samples of lower than 0.5 are classified as a low-scoring group. In the unfair dataset, α represents the degree of unfairness. As α increases, the proportion of *priv.* in the high-scoring group increases, and the proportion of *unpriv.* in the low-scoring group increases. That is, the dataset becomes more unfair. We conducted experiments with datasets with α values of 2, 3, and 4.

HR dataset	Original		$\alpha = 2$		$\alpha = 3$		$\alpha = 4$	
Demographic group	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>
High-scoring ($score \geq 0.5$)	800 (50%)	800 (50%)	800 (67%)	400 (33%)	800 (75%)	263 (25%)	800 (80%)	200 (20%)
Low-scoring ($score < 0.5$)	800 (50%)	800 (50%)	400 (33%)	800 (67%)	263 (25%)	800 (75%)	200 (20%)	800 (80%)
Total	1600	1600	1200	1200	1063	1063	1000	1000

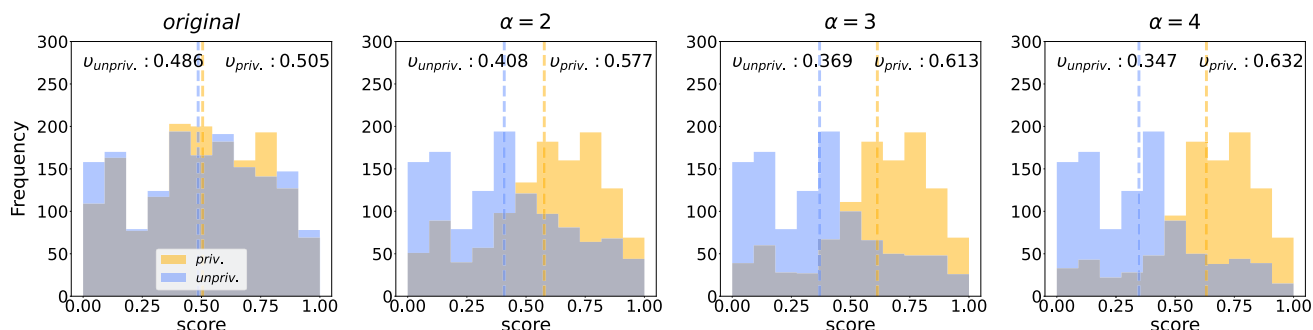


FIGURE 3. Histogram of score in HR dataset, increasing the α value made the dataset more unfair. v is the average score for each sensitive attribute S . As the value of α increases, it can be observed that the average of *unpriv.* $v_{unpriv.}$ gradually decreases and the average of *priv.* $v_{priv.}$ gradually increases. In other words, in the low-scoring group, the number of *unpriv.* is distributed relatively higher than the number of *priv.*. Whereas in the high-scoring group, the number of *unpriv.* is distributed lower than the number of *priv.*.

respect to the score group and sensitive attribute. As α increases, a significant difference can be observed in the composition ratios between the *priv.* and *unpriv.* groups in both high and low-scoring groups.

To verify that our proposed method can be applied even in unfair situations, we intentionally manipulated the original train dataset by experimenting on various settings of α . Therefore, given the uniform distribution of *score* for both sensitive attributes in the distribution of the HR and FI datasets, we used random sampling to create an unfair subset. Specifically, in the high-scoring group, the number of *unpriv.* candidates are manipulated to be smaller than the number of *priv.* candidates, while the number of *priv.* candidates in the low-scoring group are adjusted to be less than the number of *unpriv.* candidates. For example, $\alpha = 2$ means that in the high-scoring group, the number of *priv.* is twice as many as the number of *unpriv.*. Therefore, we achieve an unfair dataset by manipulating the number of *priv.* in high-scoring group and manipulating the number of *unpriv.* in low-scoring group. The experiment was conducted in three different α settings ($\alpha = 2, 3, 4$). Figures 3 and 4 show the distribution of the sensitive attribute as α changes. A significant difference is shown in the distribution of sensitive attributes between the high and low-scoring groups.

A t-test and one-way ANOVA were conducted to confirm that the distribution between sensitive attributes was different according to the α value. The results confirmed that the P-value is less than 0.05, and the distributions among the sensitive attribute groups were different.

It’s worth noting that the test dataset has an even distribution of *scores* across sensitive attributes without manipulation. It is necessary to conduct experiments under realistic conditions. On the HR dataset, we use five-fold cross-validation where 80% of the total data is utilized for training and the remaining 20% is the test dataset. For the FI dataset, holdout cross-validation is used for training.

To measure the unfairness of the dataset, we utilized modified versions of the fairness metrics DP and SPDD. In the DP and SPDD metrics, the predicted variable \hat{Y} is replaced with the target variable \tilde{Y} , referred to as the DP_{data} and $SPDD_{data}$. Table 1 shows how much bias exists in the subset of unfairly manipulated train datasets. We experimentally confirmed the extent to which the model learns manipulated train unfair subset bias and how bias affects the *score Y* prediction. By setting these unfair datasets and letting the model learn the bias in them, we can check the results thereby proving and understanding the efficiency of our proposed method. The results of this are discussed in section VI-A.

C. PRE-PROCESSING

We utilized multitask cascaded convolutional networks (MTCNN) [39] to extract the bounding boxes of faces present in the video frames. We then cropped the video frames along the bounding boxes and resized them to 112×112 . We divided the video into 30 segments of equal length and

TABLE 3. Distribution of target(score) and sensitive attributes(gender) in the FI unfair dataset (threshold = 0.5). Samples with a score of 0.5 and higher are classified as a high-scoring group, and samples of lower than 0.5 are classified as a low-scoring group. In the unfair dataset, α represents the degree of unfairness. As α increases, the proportion of *priv.* in the high-scoring group increases, and the proportion of *unpriv.* in the low-scoring group increases. That is, the dataset becomes more unfair. We conducted experiments with datasets with α values of 2, 3, and 4.

FI dataset	Original		$\alpha = 2$		$\alpha = 3$		$\alpha = 4$	
Demographic group	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>	<i>priv.</i>	<i>unpriv.</i>
High-scoring ($score \geq 0.5$)	1346 (50%)	1346 (50%)	1346 (67%)	673 (33%)	1346 (75%)	448 (25%)	1346 (80%)	336 (20%)
Low-scoring ($score < 0.5$)	1346 (50%)	1346 (50%)	673 (33%)	1346 (67%)	448 (25%)	1346 (75%)	336 (20%)	1346 (80%)
Total	2692	2692	2019	2019	1794	1794	1682	1682

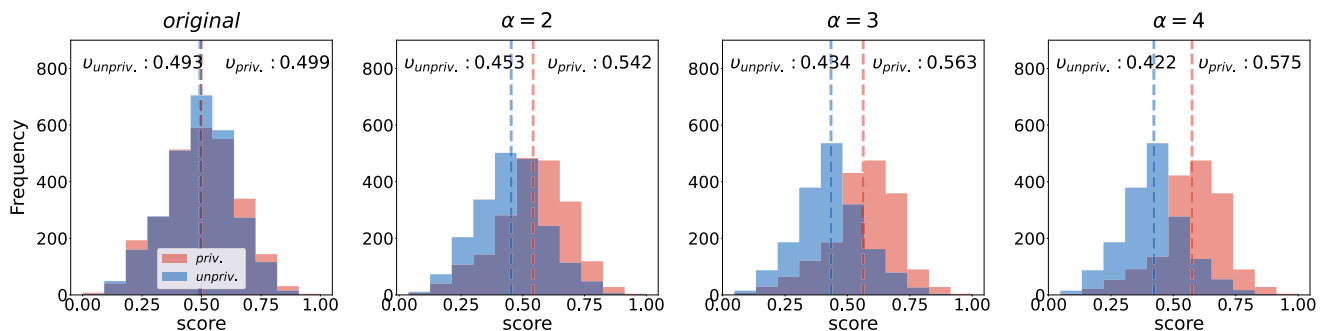


FIGURE 4. Histogram of score in FI dataset, increasing the α value made the dataset more unfair. v is the average score for each sensitive attribute S . As the value of α increases, it can be observed that the average of *unpriv.* $v_{unpriv.}$ gradually decreases and the average of *priv.* $v_{priv.}$ gradually increases. In other words, in the low-scoring group, the number of *unpriv.* is distributed relatively higher than the number of *priv.*. Whereas in the high-scoring group, the number of *unpriv.* is distributed lower than the number of *priv.*

randomly sampled one frame from each segment. As a result, 30 sampled frames were finally used as the video input X_V .

Next, we utilized an open-source library called pyAudioAnalysis [40] to extract the audio features from a raw wave file. The extracted audio features include various features including Mel Frequency Cepstral Coefficients (MFCCs). To align with the number of video frames, the resulting features have dimensions of 30×68 . These features were used as the audio input X_A . We used the same pre-processing method for both the video and audio data in both the HR and FI datasets.

Lastly, in the HR dataset, the text is obtained by utilizing the Google Speech-to-Text API,¹ which transcribes spoken words into text. The resulting text is then tokenized using a BERT tokenizer [41]. For the FI dataset, the text is obtained from the transcription CSV file provided by Chalearn Lab, which is also tokenized using the BERT tokenizer.

D. BASELINES

To demonstrate that our proposed method successfully controls the trade-off between accuracy and fairness, we compare our method with the following three models.

Vanilla: A trained model F_θ, g_ϕ without fairness constraints, using only MSE loss. The architecture of this model is identical to that of our proposed model but does not include an adversary h_ψ .

Data Balancing (DB) [25]: We utilized a resampling strategy based on the sensitive attribute and data distribution. First, we calculate the histogram of the *score* for each

sensitive attribute. Then we resample the data within the *priv.* group by down-sampling. The network architecture is the same as that of the Vanilla model.

Adversarial Learning (Adv) [42]: Trained the multimodal encoder F_θ and the regressor g_ϕ using adversarial loss:

$$L_{adv} = L_{reg} - \lambda * L_{cls}. \quad (24)$$

where L_{reg} is the MSE loss used for *score* prediction, and L_{cls} denotes the binary cross-entropy loss (BCE) used for sensitive attribute classification where λ is a hyperparameter. After the regressor g_ϕ and multimodal encoder F_θ are updated by using L_{adv} , and the adversary h_ψ is sequentially updated by using L_{cls} . We employed a similar architecture to our proposed method.

Euclidean Distance (L2): Instead of Wasserstein distance, we used Euclidean distance between *priv.* and *unpriv.* in the latent space. We employed a method of calculating the Euclidean distance between groups within the latent space and directly adding it to the loss function.

Maximum Mean Discrepancy (MMD) [43]: Domain adaptive fairness approach that tries to minimize discrepancy of fairness metrics between *priv.* and *unpriv.* domains by applying MMD loss.

E. ARCHITECTURE AND IMPLEMENTATION DETAILS

The proposed model utilizes three distinct types of inputs: video X_V , audio X_A , and text X_T . To encode the video input X_V , we stacked multiple frames as input and applied them to five 2D convolutional layers with max-pooling layers, followed by a Long Short-Term Memory (LSTM) network

¹<https://cloud.google.com/speech-to-text/docs>

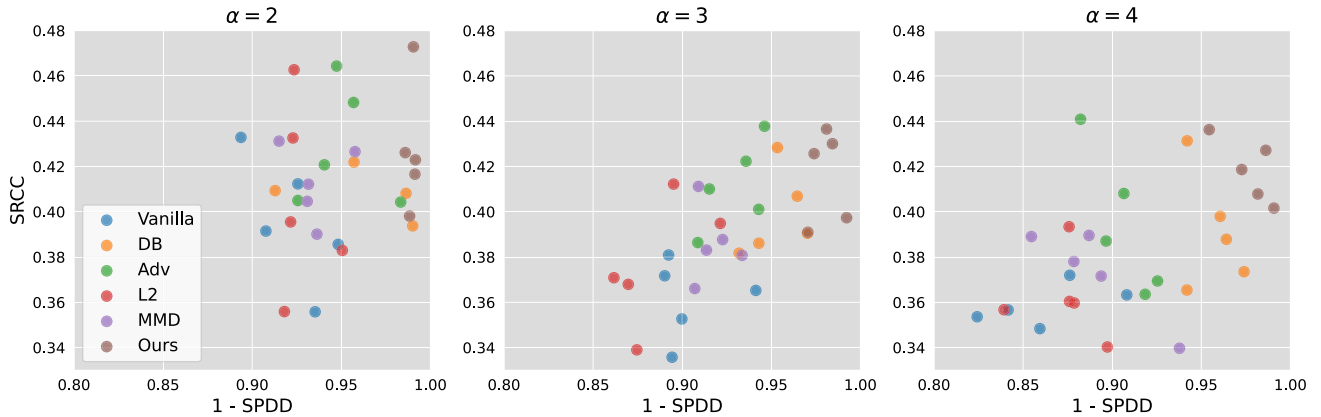


FIGURE 5. Visualization results for prediction performance (SRCC) and the fairness metric (SPDD) using the HR dataset. The x-axis is 1-SPDD. The lower the SPDD, the larger the value on the x-axis. The y-axis represents the value of SRCC. The more the result is located in the upper right corner, the more the model is considered fairer and has better performance.

TABLE 4. Results of the HR dataset. The table presents a comprehensive and comparative analysis, offering a detailed examination and evaluation of the performance outcomes achieved by the baselines and ours. By comparing respective results, the table facilitates a side-by-side comparison, shedding light on the relative strengths and weaknesses of each approach.

Method	Unfairness	PCC	SRCC	SPDD	SPEO
Vanilla	$\alpha = 2$	0.452 (0.022)	0.396 (0.026)	0.078 (0.019)	0.074 (0.021)
DB		0.466 (0.017)	0.424 (0.031)	0.037 (0.028)	0.035 (0.026)
Adv		0.471 (0.019)	0.435 (0.023)	0.057 (0.011)	0.052 (0.008)
L2		0.451 (0.028)	0.406 (0.038)	0.073 (0.012)	0.071 (0.020)
MMD		0.459 (0.018)	0.413 (0.015)	0.066 (0.014)	0.067 (0.008)
Ours		0.463 (0.022)	0.427 (0.025)	0.011 (0.002)	0.013 (0.004)
Vanilla	$\alpha = 3$	0.418 (0.012)	0.361 (0.016)	0.096 (0.019)	0.089 (0.023)
DB		0.437 (0.009)	0.399 (0.017)	0.047 (0.014)	0.043 (0.016)
Adv		0.453 (0.023)	0.412 (0.018)	0.070 (0.015)	0.063 (0.012)
L2		0.427 (0.018)	0.377 (0.025)	0.115 (0.021)	0.112 (0.014)
MMD		0.434 (0.012)	0.386 (0.015)	0.083 (0.010)	0.076 (0.007)
Ours		0.460 (0.021)	0.416 (0.018)	0.020 (0.008)	0.016 (0.006)
Vanilla	$\alpha = 4$	0.408 (0.014)	0.359 (0.008)	0.138 (0.029)	0.142 (0.029)
DB		0.434 (0.016)	0.391 (0.023)	0.043 (0.013)	0.041 (0.019)
Adv		0.433 (0.024)	0.394 (0.028)	0.094 (0.015)	0.098 (0.032)
L2		0.407 (0.015)	0.362 (0.017)	0.127 (0.019)	0.129 (0.020)
MMD		0.420 (0.012)	0.374 (0.018)	0.110 (0.027)	0.104 (0.027)
Ours		0.458 (0.006)	0.418 (0.021)	0.020 (0.008)	0.016 (0.006)

with two hidden layers f_V . For audio inputs X_A , we employed a stack of one fully connected layer and an LSTM with two hidden layers f_A . For text encoding, we utilized a pre-trained BERT [41] followed by one fully connected layer f_T . The encoded feature from each input modality was then concatenated and passed to both the adversary h_ψ and the score regressor g_ϕ . The score regressor and adversary were composed of two fully connected layers with one ReLU activation function.

For the HR dataset, we trained our model and the baselines with a batch size $m = 128$. To prevent overfitting, we set different learning rates for the regressor g_ϕ and adversary h_ψ .

It is because adversary training is relatively easier compared to regressor training. We used the AdamW [44] optimizer with a learning rate $\eta_{reg} = 10^{-3}$ for the regressor g_ϕ and learning rate $\eta_{cls} = 10^{-4}$ for the adversary h_ψ . For the FI dataset, we set a batch size $m = 64$ and optimized our model and the baselines using AdamW with a learning rate $\eta_{reg} = 10^{-4}$ and $\eta_{cls} = 10^{-5}$. In all datasets, we trained our model and the baselines for 200 epochs.

In accordance with the experience gained from experimenting with various λ_W values, the hyperparameter λ_W was set as 0.02. This setting showed the least decrease in prediction performance and effectively improved the fairness

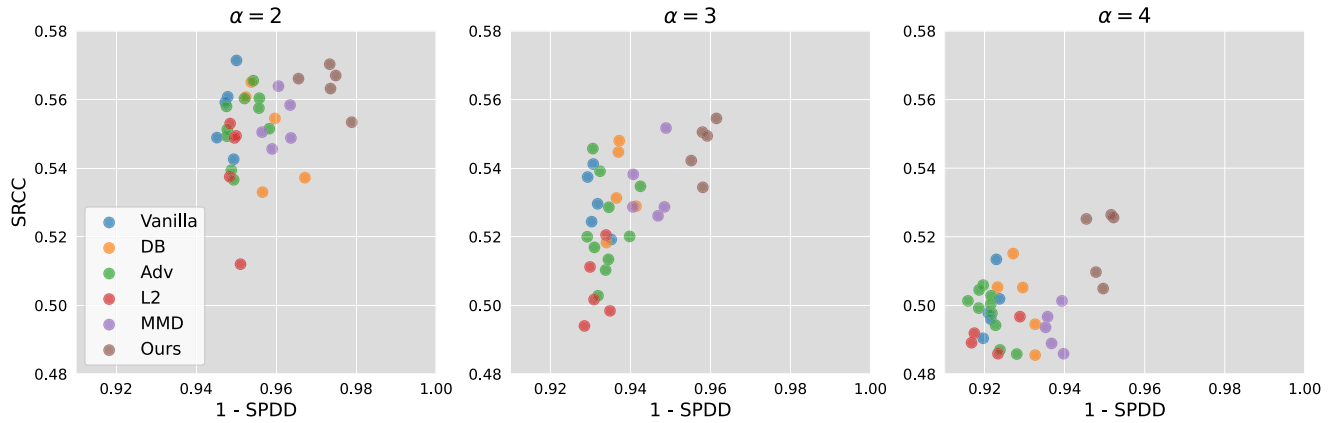


FIGURE 6. Visualization results for prediction performance(SRCC) and the fairness metric(SPDD) using the FI dataset. The x-axis is 1-SPDD, and the lower the SPDD, the larger the value on the x-axis. The y-axis represents the value of SRCC. The higher the result is located in the upper right corner, the more fair and better the performance is.

TABLE 5. Results from the FI dataset. The table presents a comprehensive comparative analysis, offering a detailed examination and evaluation of the performance outcomes achieved by both the baselines and ours. By comparing respective results, the table facilitates a side-by-side comparison, shedding light on the relative strengths and weaknesses of each approach.

Method	Unfairness	PCC	SRCC	SPDD	SPEO
Vanilla	$\alpha = 2$	0.577 (0.011)	0.557 (0.010)	0.052 (0.002)	0.054 (0.003)
DB		0.572 (0.012)	0.550 (0.013)	0.042 (0.005)	0.041 (0.009)
Adv		0.572 (0.005)	0.549 (0.007)	0.050 (0.004)	0.049 (0.006)
L2		0.563 (0.010)	0.540 (0.015)	0.050 (0.001)	0.052 (0.037)
MMD		0.577 (0.006)	0.553 (0.007)	0.039 (0.003)	0.037 (0.005)
Ours		0.582 (0.006)	0.564 (0.006)	0.027 (0.004)	0.022 (0.005)
Vanilla	$\alpha = 3$	0.552 (0.009)	0.530 (0.008)	0.069 (0.002)	0.074 (0.004)
DB		0.557 (0.009)	0.534 (0.011)	0.063 (0.002)	0.069 (0.004)
Adv		0.548 (0.010)	0.530 (0.011)	0.068 (0.002)	0.075 (0.005)
L2		0.529 (0.009)	0.505 (0.010)	0.068 (0.002)	0.076 (0.003)
MMD		0.556 (0.009)	0.535 (0.009)	0.055 (0.004)	0.057 (0.004)
Ours		0.565 (0.012)	0.546 (0.007)	0.042 (0.002)	0.044 (0.004)
Vanilla	$\alpha = 4$	0.525 (0.005)	0.500 (0.008)	0.078 (0.001)	0.087 (0.003)
DB		0.522 (0.008)	0.501 (0.010)	0.071 (0.004)	0.075 (0.005)
Adv		0.522 (0.009)	0.498 (0.006)	0.080 (0.003)	0.086 (0.005)
L2		0.513 (0.006)	0.488 (0.008)	0.079 (0.005)	0.087 (0.008)
MMD		0.520 (0.003)	0.493 (0.005)	0.063 (0.002)	0.067 (0.003)
Ours		0.540 (0.007)	0.518 (0.009)	0.051 (0.002)	0.055 (0.005)

metrics among the settings we experimented with. We also removed the information about the sensitive attribute by using adversarial learning. The adversarial hyperparameter λ_{Adv} is chosen among various values between 10^{-3} and 0.5 on a logarithmic scale.

F. EVALUATION METRICS AND PROTOCOLS

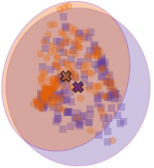
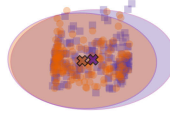
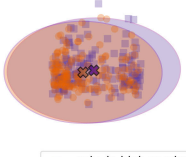
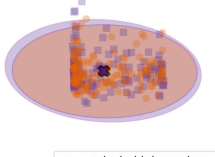
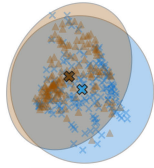
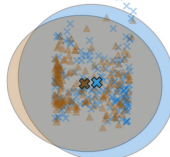
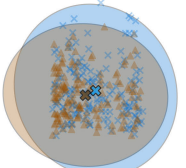
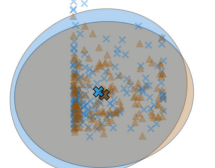
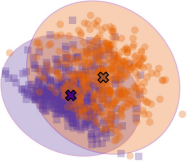
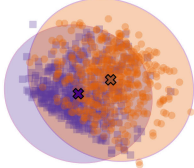
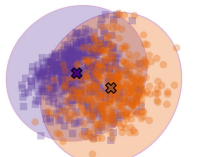
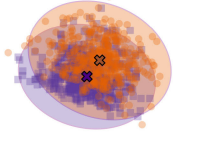
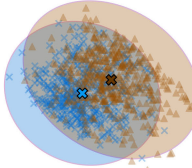
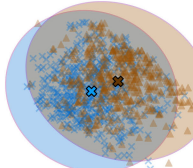
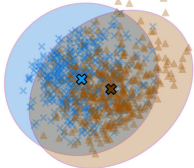
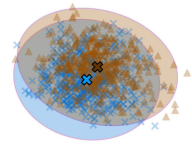
To evaluate the regression performance of predicting scores, we employed two correlation coefficients: the Pearson correlation coefficient (PCC) and Spearman’s rank correlation coefficient (SRCC). The PCC measures the degree of

association between two continuous variables, defined as:

$$PCC(\tilde{Y}, \hat{Y}) = \frac{cov(\tilde{Y}, \hat{Y})}{\sigma_{\tilde{Y}}\sigma_{\hat{Y}}}. \tag{25}$$

where n represents the number of data points, x indicates the predictions, \tilde{y} indicates the target values, and \bar{x} and \bar{y} represent the averages of the predictions, and the ground-truth values, respectively. The range of PCC is between -1 to 1 , where -1 indicates a perfect negative linear correlation, 1 indicates a perfect positive linear correlation, and 0 indicates no linear correlation between variables. In addition to the PCC, we utilized SRCC as an additional measure of regression performance. SRCC measures sta-

TABLE 6. PCA visualization of latent representations from baselines and ours using the HR and FI dataset; The center points of each group are calculated as the average of the latent representations of *priv.* and *unpriv.* samples within the high-scoring and low-scoring groups. Purple and orange circles indicate the regions where *priv.* and *unpriv.* samples within the high-scoring group are located in the latent space. While blue and brown circles indicate the regions where *priv.* and *unpriv.* samples within the low-scoring group are located in the latent space.

Dataset	Group	Vanilla	DB	Adv	Ours
HR	High scoring				
	Low scoring				
FI	High scoring				
	Low scoring				

tistical dependence between the ranking of two variables, capturing both linear and non-linear relationships. SRCC is defined as:

$$SRCC(\tilde{Y}, \hat{Y}) = PCC(R(\tilde{Y}), R(\hat{Y})). \quad (26)$$

where R represents a ranking variable.

To measure the fairness of our model, we used the fairness metric SPDD. Furthermore, in our settings, we present a fairness metric called SPEO (Strong Pairwise Equalized Opportunity), which considers the true label similar to the EO. Given a classifier g_τ , the EO gap with respect to sensitive attribute S is defined as

$$\Delta EO(g_\tau) = \mathbb{E}_{v \sim U([0,1])} |\Pr(\hat{Y} = 1 | S = 1, \tilde{Y} > v) - \Pr(\hat{Y} = 1 | S = 0, \tilde{Y} > v)|. \quad (27)$$

Consequently, the SPEO can be defined as

$$SPEO(\eta) = \mathbb{E}_{\tau \sim U([0,1])} \Delta EO(g_\tau), \quad (28)$$

where U denotes uniform distribution and η is a score function. We include SPEO as an additional fairness metric to double-check that we are conducting fair training.

In the context of fairness learning, there exists a trade-off between fairness metrics and the metric of the target task. For instance, the model may obtain a perfectly fair prediction even if the prediction of the target task is incorrect. For this reason, we selected the optimal model through cross-validation.

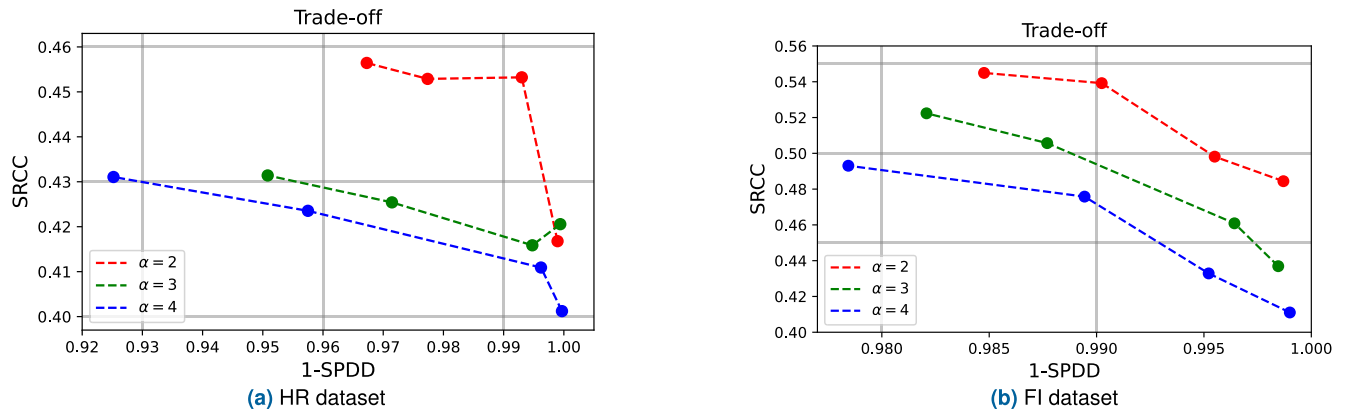


FIGURE 7. Trade-off scatter plot; The plots demonstrate the results obtained by varying the value of the λ_W . By setting a larger λ_W value, we can find that the model shows more fair results, while a smaller λ_W value improves prediction performance. The λ_W values were set within the range of 0.03 to 0.5.

VI. EXPERIMENTAL RESULTS

A. COMPARISON WITH BASELINES

Table 4 presents the prediction performance of our model and the baselines using the HR dataset, in terms of both the PCC and SRCC for score prediction. It also includes fairness metrics such as SPDD and SPEO. We report the average scores and standard deviations across five different cross-validation folds. Figure 5 is a scatter plot visualization of SRCC and 1-SPDD for the HR dataset. Each point indicates the performance of the model trained with a particular fold.

Table 4 and Figure 5 show that our proposed method significantly outperforms all three baselines in terms of both prediction correlations and fairness on the HR dataset in various degrees of unfairness (represented by α) settings. As the α value in the dataset used for model training increases, the baselines show a decline in terms of both prediction performance (SRCC) and an increase the fairness metrics (SPDD). In contrast, **Ours** shows only a slight decrease in both SRCC (less than 0.02) and a slight increase in fairness metric (less than 0.01). Remarkably, we observe that **Ours** consistently maintains low fairness scores even when the dataset is extremely unfair ($\alpha = 4$). Note that when $\alpha = 4$ in the HR dataset, the total number of data becomes 1000, which can be considered a small dataset. As a result, we can conclude that our methods work well in small datasets. While **DB** and **Adv** also show better performance than a **Vanilla** model, **Ours** surpasses them by a considerable margin in terms of both SRCC and SPDD.

Additionally, it is well-known that a trade-off exists between fairness metrics and prediction performance when applying fairness-aware training. Interestingly, **Ours** not only achieves performance gains in fairness metrics but also demonstrates higher PCC and SRCC compared to the baselines, which contradicts conventional wisdom.

To demonstrate the applicability of our proposed method to various datasets, we conducted experiments on the FI dataset and summarized the results in Table 5 and Figure 6. Similar to the results on the HR dataset, **Ours** outperforms the baselines

in terms of fairness metrics. While the baselines show a decrease in terms of fairness metrics (i.e., SPDD and SPEO increase) as α increases, **Ours** successfully maintains low SPDD and SPEO values. Additionally, it shows high PCC and SRCC values compared to the baselines.

As the value of α increases, both our model and the baselines exhibit a deterioration in prediction performance and fairness metrics. This indicates that if the dataset contains a significant bias, the model learns and predicts the bias, resulting in more unfair outcomes. As shown in Table 1, the SPDD value of the FI dataset with an α of 4 is 0.076. The model trained on this dataset ($\alpha = 4$) and tested with a uniform test dataset yielded an SPDD value of 0.078, suggesting that the model has learned and retained the bias.

In summary, our findings can be summarized as follows: when fairness considerations are not taken into account in the training phase, the extreme unfairness of the dataset may result in biased predictions by the model. In contrast, our method demonstrates the capability to mitigate extreme unfairness while simultaneously maintaining reasonable levels of prediction correlations.

B. VISUALIZATION IN LATENT SPACE

To verify whether our model has learned a fair representation, we visualized the concatenated representation Z_C obtained through a multimodal encoder F_θ . To visually confirm the representation, we perform dimensionality reduction using Principle Component Analysis (PCA) to reduce the dimensionality to 2. This allows us to visualize the representation in a two-dimensional space.

Table 6 shows the visualization of latent representation, produced by the baselines and ours from the HR and the FI datasets. As observed in the **Vanilla** representation, there is a clear distinction between the sensitive attributes. This means that the distribution of latent representation Z_C varies depending on the sensitive attribute S . Similarly to the **Vanilla**, the representations obtained through **Adv** and **DB** exhibit a clear distinction between the sensitive attribute.

TABLE 7. Ablation study: To investigate the impact of individual components on both the model's performance and fairness metrics, we conducted an ablation study by incrementally adding each component.

Method	Unfairness	PCC		SRCC		SPDD		SPEO	
		HR	FI	HR	FI	HR	FI	HR	FI
Vanilla	$\alpha = 2$	0.452 (0.022)	0.577 (0.011)	0.396 (0.026)	0.557 (0.010)	0.078 (0.019)	0.052 (0.002)	0.074 (0.021)	0.054 (0.003)
WD		0.484 (0.021)	0.577 (0.008)	0.440 (0.030)	0.553 (0.009)	0.018 (0.013)	0.037 (0.003)	0.017 (0.012)	0.037 (0.005)
Ours		0.463 (0.022)	0.582 (0.006)	0.427 (0.025)	0.564 (0.006)	0.013 (0.004)	0.027 (0.004)	0.013 (0.004)	0.022 (0.005)
Vanilla	$\alpha = 3$	0.418 (0.012)	0.552 (0.009)	0.361 (0.016)	0.530 (0.008)	0.096 (0.019)	0.069 (0.002)	0.089 (0.023)	0.074 (0.004)
WD		0.468 (0.016)	0.555 (0.013)	0.420 (0.026)	0.536 (0.009)	0.026 (0.009)	0.048 (0.004)	0.022 (0.007)	0.049 (0.006)
Ours		0.460 (0.021)	0.565 (0.007)	0.416 (0.018)	0.546 (0.007)	0.020 (0.008)	0.042 (0.002)	0.016 (0.006)	0.044 (0.004)
Vanilla	$\alpha = 4$	0.408 (0.014)	0.525 (0.005)	0.359 (0.008)	0.500 (0.008)	0.138 (0.029)	0.078 (0.001)	0.142 (0.029)	0.087 (0.003)
WD		0.463 (0.022)	0.530 (0.008)	0.422 (0.025)	0.503 (0.007)	0.041 (0.014)	0.056 (0.003)	0.036 (0.016)	0.061 (0.003)
Ours		0.458 (0.006)	0.540 (0.007)	0.418 (0.021)	0.518 (0.009)	0.020 (0.008)	0.051 (0.002)	0.016 (0.006)	0.055 (0.005)

TABLE 8. Additional experiments: In order to validate the efficacy of our approach across diverse datasets, we conducted additional experiments using the CelebA single-modal dataset and the adult tabular dataset.

Datasets	CelebA				Adult			
	gender		age		gender		race	
	Accuracy	SPDD	Accuracy	SPDD	Accuracy	SPDD	Accuracy	SPDD
Vanilla	0.810 (0.002)	0.397 (0.014)	0.807 (0.002)	0.402 (0.013)	0.854 (0.001)	0.197 (0.004)	0.853 (0.001)	0.107 (0.003)
DB	0.806 (0.003)	0.383 (0.028)	0.797 (0.004)	0.400 (0.015)	0.852 (0.001)	0.200 (0.002)	0.849 (0.001)	0.107 (0.002)
ADV	0.809 (0.003)	0.395 (0.015)	0.806 (0.002)	0.407 (0.005)	0.855 (0.001)	0.199 (0.002)	0.853 (0.001)	0.104 (0.002)
L2	0.808 (0.002)	0.367 (0.018)	0.808 (0.003)	0.402 (0.01)	0.853 (0.001)	0.174 (0.004)	0.854 (0.001)	0.086 (0.004)
MMD	0.761 (0.005)	0.138 (0.020)	0.771 (0.008)	0.224 (0.021)	0.849 (0.001)	0.085 (0.004)	0.853 (0.001)	0.040 (0.003)
Ours	0.755 (0.013)	0.089 (0.038)	0.772 (0.006)	0.202 (0.047)	0.845 (0.001)	0.071 (0.002)	0.854 (0.001)	0.031 (0.005)

However, **Ours** does not exhibit a clear visual distinction, as we have obtained an independent representation Z_C with respect to the sensitive attribute.

We have not only visually demonstrated these findings but also quantitatively measured them. We calculated the centroid (ν) within the high and low-scoring groups based on the sensitive attribute. Indeed, if the distance between the centroid of priv. representations and the centroid of unpriv. representations are small, it can be considered a fair representation. **Ours** demonstrated the smallest distance within the high-scoring and low-scoring groups. This can be interpreted as the distribution of representation Z_C is relatively independent of the sensitive attribute.

C. FAIRNESS AND CORRELATION TRADE-OFF

It is known that there is a trade-off between prediction performance and fairness metrics. Our proposed method allows users to adjust this trade-off by modifying the weights of the loss function.

Figure 7 represents the experimental results obtained by varying the values of λ_W and measuring the SPDD and SRCC in each λ_W . We assessed the results after conducting five repeated experiments for each combination of various unfair situations and different λ_W values. The results show that as the value of λ_W increases, the fairness, represented by 1-SPDD, also increases. Conversely, as the λ_W decreases, the SRCC tends to increase. By adjusting the λ_W parameter, it becomes possible to address the trade-off between prediction performance and fairness metrics, as demonstrated in this method.

D. ABLATION STUDY

To assess the significance of specific architectural components within the model, we conducted an ablation study. Table 7 presents the results of the ablation study conducted on the HR and FI datasets. First, results of **Vanilla** model without any regularizer are shown. Subsequently, the outcomes of only utilizing the Wasserstein Distance (**WD**) as a regularizer on the sensitive attribute as gender are demonstrated, followed by the results of incorporating domain adaptation (**Ours**) to eliminate gender-related information. The results of HR dataset exhibit notable high values for PCC and SRCC in **WD**-only trained results. On the other hand, on FI dataset, our proposed approach demonstrates superior performance in terms of PCC and SRCC. In both datasets, our method shows better performance in terms of SPDD and SPEO.

E. ADDITIONAL EXPERIMENTS

To verify the robustness of our approach across various datasets and different sensitive attributes, we conducted experiments by introducing two additional datasets and two distinct sensitive attributes. The utilized datasets for the experiments include the CelebA dataset [45] and the Adult Income Dataset [46]. This selection aims to ensure the efficacy of our approach across both single-modal and tabular data domains. The CelebA dataset contains 202,599 face images, each of resolution 178×218 , with 40 binary attributes. We consider binary classification tasks with two different sensitive attributes: predicting attractiveness with gender as a sensitive attribute and predicting attractiveness with age as a sensitive attribute. The Adult income dataset

contains 65,123 samples with 14 attributes and one binary label indicating if an individual's income exceeds 50K. For the Adult income dataset, the goal is to predict whether an individual's income exceeds 50K, and the sensitive attributes are gender and race.

Table 8 presents the results obtained from experiments conducted using the CelebA and Adult income datasets. Upon comparing with alternative baselines, our method demonstrated the most favorable outcomes in terms of the fairness metric, specifically SPDD, while showing decent performance in accuracy. We can also investigate that the SPDD is high when we use age as a sensitive attribute. We assume that this is because the attractiveness is highly correlated with age. Nevertheless, our method has shown the best results compared to baseline methods.

VII. DISCUSSION AND CONCLUSION

In this paper, we present a novel approach for developing a fair and accurate automatic interview assessment model. Our proposed method involves minimizing the 1-Wasserstein distance between predicted scores of different groups and learning fair multimodal representations by leveraging gradient reversal layers to dilute group information defined by sensitive attributes. Our experiments show that our proposed method outperforms existing methods in terms of accuracy and fairness criteria, such as SPDD and SPEO.

To the best of our knowledge, our work is first to focus on fairness in the context of automatic interview assessment that involves multimodal features. Although previous studies have explored fair automatic recruitment, they have mostly been conducted on tabular or synthetic datasets. Furthermore, we conducted extensive experiments on a real-world job interview dataset called the HR dataset, and a public benchmark dataset called the FI dataset. We also found that our proposed method is robust to the degree of unfairness in the training dataset, which is a crucial factor for the practical applicability of the method.

However, all of these solutions have a limitation in that sensitive attributes need to be labeled. In recent privacy-preserving environments, there exist datasets that do not have labeled sensitive attributes. Due to the enactment of the E.U. general data protection regulation, data protection laws have been strengthened, making it more challenging to collect personal information such as age, gender, and race. Even in the absence of that information, it is necessary to develop fair models. We consider it as future work. Also, we tested our method considering only the most commonly used AVI method. Therefore, we plan to conduct further research by considering a wider range of interview methods in future studies and see how they affect fairness. Moreover, there exists a research area regarding fair sharing that aims to distribute goods fairly among users or groups. We plan to work with various perspectives in fairness in the future [47], [48], [49].

REFERENCES

- [1] 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," *Hum. Resource Manage. Rev.*, vol. 32, no. 1, Mar. 2022, Art. no. 100789.
- [2] N. Takeuchi and T. Koda, "Job interview training system using multimodal behavior analysis," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2021, pp. 1–3.
- [3] Y.-C. Chou, F. R. Wongso, C.-Y. Chao, and H.-Y. Yu, "An AI mock-interview platform for interview performance analysis," in *Proc. 10th Int. Conf. Inf. Educ. Technol. (ICIET)*, Apr. 2022, pp. 37–41.
- [4] S. Anglekar, U. Chaudhari, A. Chitanvis, and R. Shankarmani, "A deep learning based self-assessment tool for personality traits and interview preparations," in *Proc. Int. Conf. Commun. Inf. Comput. Technol. (ICCICT)*, Jun. 2021, pp. 1–3.
- [5] L. Hickman, N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo, "Automated video interview personality assessments: Reliability, validity, and generalizability investigations," *J. Appl. Psychol.*, vol. 107, no. 8, p. 1323, 2022.
- [6] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of Data and Analytics*. Boca Raton, FL, USA: Auerbach Publications, 2018, pp. 296–299.
- [7] D. Pessach and E. Shmueli, "Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115667.
- [8] A. Peña, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal AI: Testbed for fair automatic recruitment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 129–137.
- [9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 259–268.
- [10] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. Artif. Intell. Statist.*, 2017, pp. 962–970.
- [11] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. Adv. Neural Inf. Process. Syst.*, 30, 2017.
- [12] C. Villani and C. Villani, "The Wasserstein distances," in *Optimal Transport: Old and New*, 2009, pp. 93–111.
- [13] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*.
- [14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
- [15] 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. Multimedia*, vol. 16, no. 4, pp. 1018–1031, Jun. 2014.
- [16] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–6.
- [17] A. L. Hunkenschroer and A. Kriebitz, "Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring," *AI Ethics*, vol. 3, no. 1, pp. 199–213, Feb. 2023.
- [18] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Mining Knowl. Discovery*, vol. 21, no. 2, pp. 277–292, Sep. 2010.
- [19] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, Oct. 2012.
- [20] B. T. Luong, S. Ruggieri, and F. Turini, "K-NN as an implementation of situation testing for discrimination discovery and prevention," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 502–510.
- [21] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowl. Inf. Syst.*, vol. 54, no. 1, pp. 95–122, Jan. 2018.
- [22] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 702–712.
- [23] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3384–3393.

- [24] E. Krasnakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 853–862.
- [25] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 361–369.
- [26] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Dec. 2018, pp. 335–340.
- [27] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 319–328.
- [28] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.
- [29] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018.
- [30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.
- [31] V. Iosifidis, B. Fetahu, and E. Ntoutsi, "FAE: A fairness-aware ensemble framework," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 1375–1380.
- [32] W. Lee, H. Ko, J. Byun, T. Yoon, and J. Lee, "Fair clustering with fair correspondence distribution," *Inf. Sci.*, vol. 581, pp. 155–178, Dec. 2021.
- [33] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Proc. Uncertainty Artif. Intell.*, 2020, pp. 862–872.
- [34] T. Yoon, J. Lee, and W. Lee, "Joint transfer of model knowledge and fairness over domains using Wasserstein distance," *IEEE Access*, vol. 8, pp. 123783–123798, 2020.
- [35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [36] Y. Ganin, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.
- [37] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Scale Space and Variational Methods in Computer Vision*. Ein-Gedi, Israel: Springer, May 2011, pp. 435–446.
- [38] H. Jair Escalante, H. Kaya, A. Ali Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baro, I. Guyon, J. Jacques Junior, M. Madadi, S. Ayache, E. Viegas, F. Gurpinar, 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," 2018, *arXiv:1802.00745*.
- [39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [40] T. Giannakopoulos, "PyAudioAnalysis: An open-source Python library for audio signal analysis," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144610.
- [41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [42] A. Bose and W. Hamilton, "Compositional fairness constraints for graph embeddings," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 715–724.
- [43] C. Schumann, X. Wang, A. Beutel, J. Chen, H. Qian, and E. H. Chi, "Transfer of machine learning fairness across domains," 2019, *arXiv:1906.09688*.
- [44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [45] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
- [46] A. Asuncion and D. Newman, "UCI machine learning repository," Tech. Rep., 2007.
- [47] F. J. Gutierrez, E. Varvarigos, and S. Vassiliadis, "Multicost routing in max-min fair share networks," in *Proc. 3rd Int. Conf. Scale Space Variational Methods Comput. Vis.*, vol. 2, Oct. 2000, pp. 1294–1304.
- [48] H. Aziz, H. Chan, and B. Li, "Weighted maxmin fair share allocation of indivisible chores," 2019, *arXiv:1906.07602*.
- [49] N. Doulamis, E. Varvarigos, and T. Varvarigou, "Fair scheduling algorithms in grids," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 11, pp. 1630–1648, Nov. 2007.



CHANGWOO KIM received the B.S. degree in computer science from Chosun University, in 2021. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Dongguk University, Seoul, South Korea. His research interests include multimodal, fairness, and time-series analysis.



JINHO CHOI received the B.S. degree from the Department of Computer Science, Korea University, in 2017, and the M.S. degree from the Department of Computer Science, Korea University, in 2019, supervised by Prof. Jaegul Choo. He is currently pursuing the Ph.D. degree with the Graduate School of AI, KAIST. He is also a AI Scientist with the Genesis Laboratory. His research interest include fairness and debiasing in deep learning.



JONGYEON YOON is currently pursuing the B.S. degree with the Department of Statistics, Dongguk University. He is also a Research Student with the Data Analysis and Machine Intelligence Laboratory, Dongguk University. His current research interests include machine learning, deep learning, fairness, and medical imaging.



DAEHUN YOO received the B.S. degree in mathematics and the master's and Ph.D. degrees in computer science from Kwangwoon University, Seoul, South Korea, in 2005, 2007, and 2013, respectively. He has been a Chief Artificial Intelligence Officer (CAIO) with the Genesis Laboratory, since 2017. His research interests include fairness and multi-modal deep learning.



WOOJIN LEE received the B.S. degree in information and industrial engineering from Yonsei University, Seoul, Republic of Korea, in 2015, and the Ph.D. degree in industrial engineering from Seoul National University, in 2020. He is currently an Assistant Professor with the College of AI Convergence, Dongguk University, Seoul. His research interests include robustness and fairness in deep learning.

...