# HirePreter: A Framework for Providing Fine-grained Interpretation for Automated Job Interview Analysis

Wasifur Rahman[1*], Sazan Mahbub[2*], Asif Salekin[3], Md Kamrul Hasan[1], Ehsan Hoque[1]
1 - Department of Computer Science, University of Rochester, USA
2 - Bangladesh University of Engineering and Technology, Bangladesh; 3 - Syracuse University, USA
echowdh2@ur.rochester.edu, 1505020.sm@ugrad.cse.buet.ac.bd, asalekin@syr.edu,
mhasan8@cs.rochester.edu, mehoque@cs.rochester.edu

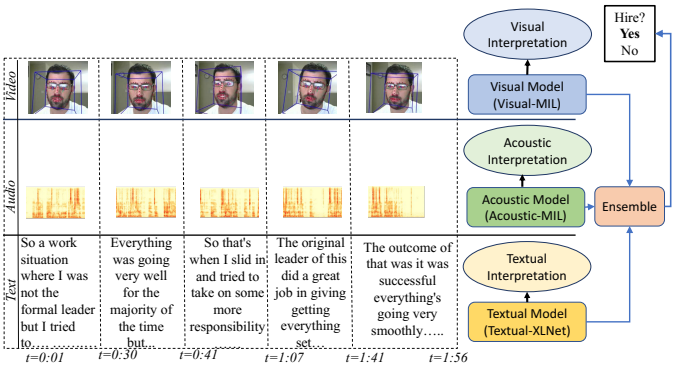



Fig. 1. A overview of our entire framework. Each video data is composed of three modalities: textual, acoustic and visual. For each modality, we train a model and build an interpretation module. The outputs of all three models are aggregated in a Ensemble manner to produce the final hiring decision.

***Abstract*—There has been a rise in automated technologies to screen potential job applicants through affective signals captured from video-based interviews. These tools can make the interview process scalable and objective, but they often provide little to no information of how the machine learning model is making crucial decisions that impacts the livelihood of thousands of people. We built an ensemble model – by combining Multiple-Instance-Learning and Language-Modeling based models – that can predict whether an interviewee should be hired or not. Using both model-specific and model-agnostic interpretation techniques, we can decipher the most informative time-segments and features driving the model's decision making. Our analysis also shows that our models are significantly impacted by the beginning and ending portions of the video. Our model achieves 75.3% accuracy in predicting whether an interviewee should be hired on the ETS Job Interview dataset. Our approach can be extended to interpret other video-based affective computing tasks like analyzing sentiment, measuring credibility, or coaching individuals to collaborate more effectively in a team.**

***Index Terms*—Interpretability, Job Interview Analysis, Fairness in AI**

## I. Introduction

Imagine a computer algorithm that conducts job interviews and weeds out candidates for the final round without any explanations. While this may sound far fetched, more than 700 companies have already conducted over 10 million such interviews using products such as HireVue [1] to quickly shortlist interviewees from a large pool of candidates by utilizing black-box proprietary algorithms. These technologies are making the interview process more scalable – adding revenue for the companies who sell these technologies as well as for the companies who buy them. However, these technologies can potentially do a fundamental disservice to the general population, especially to individuals from lower socioeconomic status and people of color [2], by making important hiring decisions with no accountability or explanations. Therefore, it is crucial that we add checks and balances to impose interpretability and transparency to avoid unintended consequences on humans, especially ones that exacerbate existing disparities in the job market against women [3], older workers [4] and minorities [5]. AI can potentially eliminate unconscious human bias through careful design choices – like increasing model's interpretability [6] – and assess a large pool of candidates fairly quickly without resorting to biased shortlisting procedures [7].

In this paper, we focus on adding interpretability in automated human behavior analysis in the context of job interviews. In particular, we present a computational framework that can automatically analyze the recorded videos (i.e., text representing content, acoustic and visual features) of job interviews, provide an outcome, and generate interpretable feedback by analyzing algorithm's decision-making process. We model the text with the Transformer-based ALBERT [8]. In addition, we use two separate Attention-based Multiple-Instance-Learning(MIL) [9] models for the acoustic and visual modalities separately. The final prediction is done by combining the decision from all three models through a neural-network. For interpreting all three models, we use the model-agnostic interpretation providers like SHAP (SHapley Additive exPlanations) [10] – a framework to understand how each feature impacts the model's outcome – to understand the salient time-segments and features in those segments that drive a model's decision. We provide an example from our Interpretation framework in Fig 2.

We use the ETS Job Interview dataset [11] consisting of 1891 monologue job interview videos (63 hours in duration) from 260 online workers from the USA for training our

This work was supported by NSF grant IIS-1750380 and W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency and the Army Research Office.

* - Equal contribution

978-1-6654-0021-3/21/$31.00 ©2021 IEEE

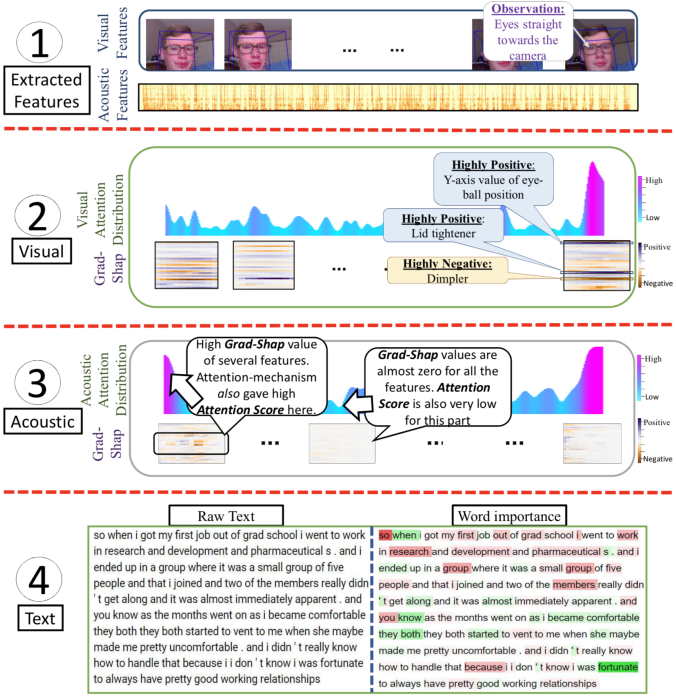



Fig. 2. Interpretation for an example video. Here the segment (1) represents extracted visual (top) and acoustic (bottom) features. The top portions of both (2) and (3) represent how visual and acoustic models put attention on a portion of the video. Similarly, the bottom portions of (2) and (3) represents feature importance (through Grad-Shap) for each feature for some selected instances. Segment (4) shows the textual interpretation, where we have raw text transcripts on the left and word importance of the text (according to layer-integrated gradients) on the right. For word importance, the color green represents that the word is likely to increase the hiring probability, and red means the opposite.

models and reporting their performance. The dataset represents a balanced mix of individuals across different ages, ethnicities and socio-economic statuses, and should exhibit real-world variability.

We summarize our contributions as:

- We built a framework that provides interpretation (Fig. 2) of the textual, acoustic and visual modalities separately for a deep-learning-based ensemble model trained to automate interview judgement (Sections III and II).
- Our model achieves performance of 75.3% accuracy on the ETS dataset [11] – an 8% increase from the previous baseline (Section IV)
- We validate our interpretation framework through several experiments (V-A). We show that our interpretation system can capture the most informative time-segments and features present in our data (V-B).

## II. Experimental Setup

In this section, we provide a short description of ETS Dataset, Feature Extraction and Training strategies.

### A. Dataset

We used the ETS-dataset introduced in [11] – a job-interview dataset containing 1891 videos from 260 Amazon Mechanical Turk workers answering questions about how they handled unfair work distribution, showed leadership, overcame a weakness, etc. Five experts from ETS rated each video on whether they will hire the candidate on a 7 point Likert scale (1= Strongly Disagree, 7 = Strongly Agree) with 0.79 inter-rater agreement score. To train a binary classifier, we followed the protocol established in [11] to take the median of hiring scores from the entire dataset as a threshold – a video with score greater than the threshold was labeled as positive (hired).

### B. Feature Extraction

We extracted features from Textual, Acoustic and Visual modalities to analyze both the verbal and non-verbal cues present in ETS dataset.

**Textual:** We transcribed the videos using YouTube transcription API[1], followed by manual correction and manual sentence boundary detection.

**Acoustic:** After experimenting with different feature set combinations (mel spectrogram, MFCC, spectogram, filterbanks, etc.), we zeroed in on the filterbanks features – created by applying multiple band-pass filters on an input signal to separate it into multiple single-frequency sub-bands – from Shennong[2] based on highest performance on Validation set.

**Visual:** Using Openface2 [12], we extracted 49 features corresponding to the Action Unit (AU)-regression, AU-classification, head-pose and eye-related features. Then, we removed correlated features and end up using 31 features (that provide us with a model with the highest validation score).

### C. Training Strategy:

144 videos in the dataset had too poor quality to pass the automated transcription. We discarded them and split the remaining 1747 data into train (1223 videos), validation (350) and test (174). An interviewee doesn't occur in more than one set so that our models cannot boost performance by learning interviewee specific idiosyncrasies. Our data-sets are mostly balanced: the fraction of positive samples were 0.513, 0.5, 0.483 in train, validation, test sets respectively. We train the model on the train set for a chosen set of hyper-parameters, choose the best set of hyper-parameters (and thereby the best model) on the validation set, and finally report the score on the test set.

## III. Model

In this section, we will discuss modeling Textual (III-A) and Acoustic/Visual (III-B) modalities, and combining information from three unimodal models. (III-C).

### A. Textual-ALBERT

As depicted in Fig. 3, we modeled Text by fine-tuning the last layer of a pre-trained ALBERT [8] – a lightweight version of the popular BERT model [13], [14]. For a given $N$ length input consisting of tokens $[L_1, L_2, \ldots L_N]$, we append a $CLS$ token (used for generating the vector used for

[1]https://pypi.org/project/youtube-transcript-api/
[2]https://github.com/bootphon/shennong

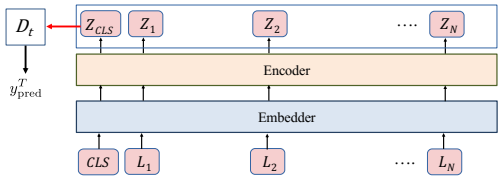



Fig. 3. Textual-ALBERT model architecture

Classification) and pass them through the *Embedder* and the *Encoder* layers. At the end, we get a sequence of self-attended vectors $[Z_{CLS}, Z_1, Z_2, \ldots Z_N]$. For the sake of consistency, we will call $Z_{CLS}$ to be $Z^T$ – representing the vector for the textual modality.

### B. Multiple Instance Learning (MIL) Model

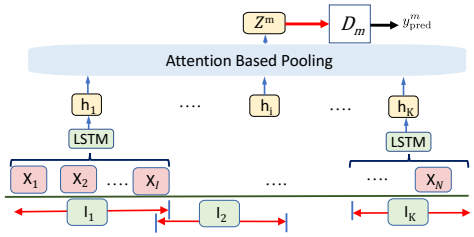



Fig. 4. Multiple Instance Learning Model architecture.

We built two separate models by modifying the MIL model [9] – one for the Acoustic and another for the Visual modalities. MIL can identify the portions of data most indicative of interview performance.

We represent an $N$ length data sequence as $X^m = [X_1, X_2, \ldots X_N]$ where $m$ represents either the Acoustic ($A$) or Visual ($V$) modality. We convert each data sequence into a bag-of-instances; the bag will contain $K$ contiguous and overlapping instances $[I_1, I_2, \ldots I_K]$. From these overlapping instances, we generate embedding vectors $[h_1, h_2, ...h_K]$, each representing their corresponding instance. Those embedding vectors are passed through an attention-based aggregation mechanism to generate a vector $Z^m$ representing the entire input data sequence. Using this generic framework, we create two separate models: *Acoustic-MIL* and *Visual-MIL*. The aggregated vectors from these two models are $Z^A$ and $Z^V$.

### C. Ensemble and Late-Fusion models

Following the description in III-A and III-B, we train three individual models: *Textual-ALBERT*, *Acoustic-MIL*, and *Visual-MIL*. For each input data, we get three corresponding vectors from the pre-trained models: $[Z_T, Z_A, Z_V]$. We concatenate these three vectors to produce a new vector $Z = Z_T \oplus Z_A \oplus Z_V$. Then, we train a neural-network unit $D$, consisting of a Linear and a Sigmoid layer, to get a final prediction: $y_{\text{pred}} = D(Z)$.

During this final joint training, we deploy two different strategies: we either keep the parameters in the three pre-trained models fixed or fine-tune them; the first strategy produces our *Ensemble* model, and the second one produces *Late-fusion* model.

## IV. Results

TABLE I
BINARY ACCURACY AND AUC FOR DIFFERENT MODELS DECRIBED IN SECTION III IN HIRING DECISION PREDICTION. BASELINE IS THE BEST MODEL FROM [11], THE PREVIOUS STATE-OF-THE-ART ON THIS DATASET.

| Models | Accuracy | AUC Score |
|---|---|---|
| Textual-ALBERT | 0.667 | 0.676 |
| Acoustic-MIL | 0.684 | 0.714 |
| Visual-MIL | 0.621 | 0.635 |
| Ensemble | **0.753** | **0.746** |
| Late-fusion | 0.724 | 0.738 |
| Baseline [3] | 0.67 | – |

Table I contains the Binary accuracy and Area-Under-the-Curve (AUC) metrics on predicting the hiring decision from all our trained models mentioned in Sec. III. Although, our *Ensemble* model outperforms the previous baseline [11], this is not a one-to-one comparison since we do not share the identical dataset (as explained in Sec. II-C).

Among the unimodal models, the Acoustic-MIL and Textual-ALBERT have similar performance and Visual-MIL performs the worst, which are in congruence with the findings in previous baseline [11]. We get a significant gain in performance by using the *Ensemble* model. This is different than the previous baseline [11], which found modeling Text modality standalone gives the best performance. Surprisingly, *Late-fusion* model does not outperform *Ensemble* model. We speculate that the *Late-fusion* model is failing to fine-tune three separate pre-trained models simultaneously without deploying advanced fusion strategies to capture cross-modality interactions.

## V. Interpretation Framework

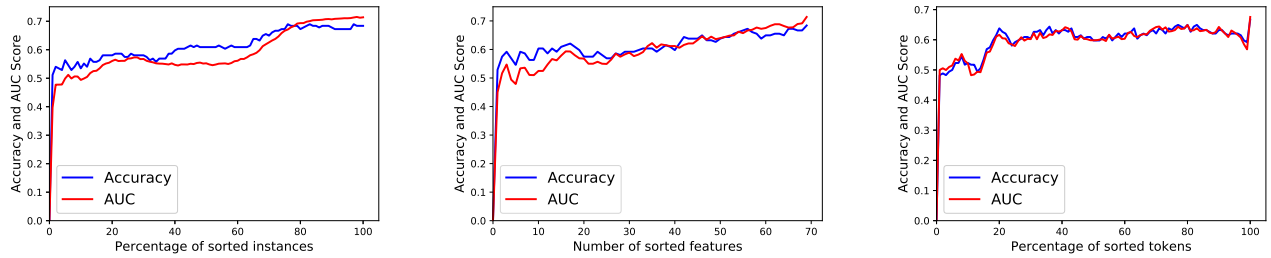
In this section, we will describe our Interpretation Framework (V-A), how we validated our approach (V-B), and which time-segments within the videos are impacting the model's decision significantly (V-C).

### A. Interpretation Technique

We use Integrated Gradient [15] on the Textual model and GradientSHAP [16] on Acoustic/Visual models. Integrated Gradients calculates each feature's attribution (denoting importance) score by integrating the gradients for the data points in between the path from a given input to a pre-defined reference [15]. GradientSHAP is a variant of SHapley Additive exPlanations (SHAP) [10] – a mathematical framework for determining the impact of each feature on the model's prediction.

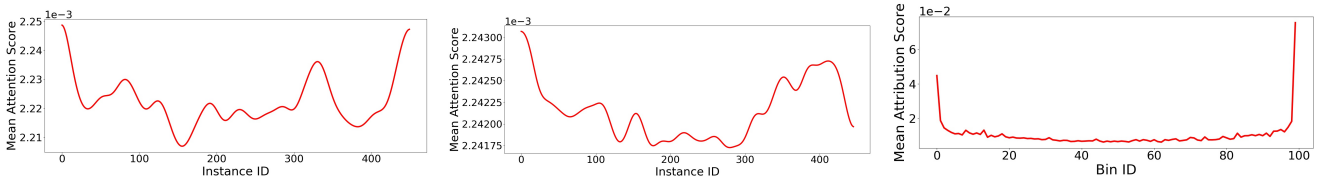
### B. Can Interpretation Framework detect the most important time-segments and features?

Our interpretation framework provides two types of feedback. First, how much attention our Acoustic/Visual models are putting on each of the instances (segments). Second, how




(a) (b) (c)

Fig. 5. Validation of (a) Attention mechanism of Acoustic-MIL. (b) Gradient-Shap values for Acoustic-MIL, (c) Attribution values for Textual-ALBERT.




(a) (b) (c)

Fig. 6. We present Mean Attention Score distribution for Visual-MIL and Attention-MIL respectively in (a) and (b), and Mean Attribution Score distribution for Textual Modality in (c).

much each token (for Text) or features (for Acoustic and Visual) are influencing the model's decision. Now, we will provide an approximate validation of our Interpretation mechanism for Acoustic instances (Fig. 5(a)), Acoustic features (Fig. 5(b)), and Textual tokens (Fig. 5(c)) across the entire Test-dataset.

In Fig. 5(a), we demonstrate how the *Acoustic-MIL* model's accuracy and AUC metrics vary with addition of the instances with the most to the least attention score. Similarly, Fig. 5(b) demonstrates the performance variation of *Acoustic-MIL* with the addition of features sorted according to Gradient-SHAP score. As both Fig. 5(a) and Fig. 5(b) clearly demonstrate, the most important segments (or features) have the most influence on the model's decision making, and with the addition of more instances (or features), the model's performance increases until a point of saturation. We can see the same pattern of performance saturation for Textual-AlBERT (in Fig 5(c)) and Visual-MIL (supplementary materials).

### C. Which parts of the data are most salient (on average) ?

Previous research on the MIT-Interview-dataset showed that models making the hiring decision are impacted greatly by the beginning and ending of a video [17]. As shown in Fig.6, that conclusion holds largely true for ETS dataset and our models as well.

As mentioned in Section III, our model learns an attention score for each of the instances in acoustic and visual modalities, and uses that scores to give appropriate importance to these instances in making the final prediction. We calculate a (normalized) mean attention score for each of these instances by averaging across all the data-points in test set. Fig. 6(a) and Fig. 6(b) show the mean attention distribution for the visual and acoustic models respectively after applying gaussian smoothing. However, since each video contains different number of words in the textual modality, we divided the words in each video in 100 different bins. Then, we calculate average attribution score in each bin, normalize the scores, apply gaussian smoothing and present the mean attribution distribution in Fig.6(c).

Although the pattern of increased importance at the beginning and end is quite apparent for the Visual and Text modalities, it is less so in the Acoustic one since there is a sharp drop of mean attention during the last 20 segments. Our assumption is: the interviewees typically stop talking few seconds before the end of allocated two minutes, and therefore, there is less data in that portion.

## VI. Conclusion

In this paper, we have built an interpretable framework for making automated hiring decision while detecting the salient time-segments and features/text-tokens driving the model's decision. Our framework significantly outperforms the baseline detects the salient time-segments and features. We also showed that our models focus on the beginning and ending portions of the videos on average. Although the interpretation techniques are not free from shortcomings, they can be a great tool to augment human reasoning with more objective, machine-derived feedback and potentially, enable a more robust human-in-the-loop hiring process in future. In future, we plan to augment our framework to enable both human-in-the-loop hiring process and give actionable and interpretable feedback to interviewees trying to hone their job interview skills.

## REFERENCES


[1] "Hirevue ·end-to-end hiring experience platform." [Online]. Available: https://hirevue.com

[2] J. Buolamwini, "When the robot doesn't see dark skin," *NY Times. Retrieved from https://www. nytimes. com/2018/06/21/opinion/facial-analysis-technology-bias. html*, 2018.

[3] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proceedings of the national academy of sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.

[4] D. Neumark, I. Burn, and P. Button, "Age discrimination and hiring of older workers," *Age*, vol. 6, 2017.

[5] M. Bertrand and S. Mullainathan, "Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination," *American economic review*, vol. 94, no. 4, pp. 991–1013, 2004.

[6] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[7] F. Polli, "Using ai to eliminate bias from hiring," *Harvard Business Review. Retrieved from https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring*, 2019.

[8] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[9] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[11] 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 *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 504–509.

[12] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[16] "Captum · model interpretability for pytorch." [Online]. Available: https://captum.ai/api/gradient$_s hap.html$

[17] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated analysis and prediction of job interview performance," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 191–204, 2016.