Counterfactual Augmentation for Speech-based Depression Detection under Data Scarcity

Lishi ZUO, *Student Member, IEEE*, Man-Wai Mak, *Senior Member, IEEE*

***Abstract*— Data scarcity is a common and serious problem in depression detection, often leading to overfitting and bias that degrade the performance of depression detectors. We propose a counterfactual augmentation (CF-aug) framework that generates latent features for speech-based depression detection under data-scarce conditions. The generation method is based on exploring how feature changes affect the outcomes. To this end, we introduce a counterfactual layer to a deep network to transform the representation of the original data to its opposite class, while a group-wise vector quantization module helps the model explore how the changes in vectors (or entries) sampled from codebooks affect the outcome. Experimental results demonstrate that CF-aug can alleviate the overfitting and bias problems caused by data scarcity. Our CF-aug framework achieves competitive performance compared to state-of-the-art methods on two depression datasets. We also demonstrate the potential of CF-aug in other domains and modalities for medical diagnosis under data-scarce settings.**

***Index Terms*— Speech-based depression detection; data scarcity; data augmentation; counterfactuals;**

I. Introduction

Depression is a significant global health issue [1]. Traditional diagnosis methods based on clinical interviews are time-consuming and subjective [2]. Speech-based automatic detection offers a promising alternative, as cognitive and physiological changes caused by depression influence speech production [3]–[5]. However, data scarcity is a common challenge in depression detection task, where obtaining large, diverse depression datasets is often hard due to privacy, stigma, and legal concerns [6]. The limited availability of annotated data can cause issues like overfitting. This is because when there is a lack of diversity in the limited training data, the models will memorize the training samples rather than generalize effectively. Moreover, the lack of data can introduce biases. For example, the work in [7] shows that a depression model can bias towards speaker features under data scarcity when speaker labels are correlated with depression labels.

An obvious solution to the problems mentioned is data augmentation. Data augmentation methods modify data to introduce variations, which can improve model performance. Variations can be introduced either by manually designed perturbations, such as flipping images or using generative adversarial networks to generate synthetic data with a distribution close to the original data [8]. In [9], augmented samples were introduced by changing the frame-width and the frame-shift of speech for depression detection. Typically, these methods generate data for different classes separately and do not consider the relationships between classes.

Understanding the relationships between different classes can be useful because it encourages the incorporation of counterfactual information [10] to medical diagnosis. From a causality perspective, counterfactuals involve asking “what-if” questions about how changes or interventions on specific features might alter outcomes. Using counterfactual information, the model can better exploit the underlying mechanisms that differentiate the positive and negative classes, leading to more accurate predictions. Importantly, this capability is crucial to medical diagnosis, where the focus often lies on the presence or absence of certain features. Understanding how these features influence diagnostic outcomes helps researchers build a model that produces interpretable decisions.

Inspired by [11], we propose a counterfactual augmentation framework (CF-aug) to address data scarcity for speech-based depression detection. The framework generates counterfactual samples to promote the model to consider the relationships between different classes. The key idea is to encourage the model to learn how the change in features can affect the prediction or outcome of the model. The core component of CF-aug is a counterfactual layer that generates latent counterfactual features by transforming the representation of the original data to their opposite class.[1] In practice, we apply group-wise vector quantization (VQ) [12] to discretize the latent features by selecting specific vectors (or entries) from multiple codebooks. Therefore, the VQ module facilitates the model to explore how changing entries in the feature groups affect the prediction outcomes and gain discriminative power. Meanwhile, the increased diversity helps the model become more robust to data variations and reduces the risk of bias and overfitting.

Contributions. We propose a novel framework to generate latent counterfactual features for speech-based depression detection under data scarcity. We evaluated CF-aug on two speech-based depression datasets and achieved state-of-the-art performance. Ablation studies were conducted to demonstrate the effectiveness of augmented counterfactual features. We demonstrate that CF-aug can reduce overfitting and bias, which are detrimental to depression detection when using small datasets. Notably, CF-aug is not limited to speech-based depression detection and can be used in other medical

The authors are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail:, lishi.zuo@connect.polyu.hk; enmwmak@polyu.edu.hk).

[1] For simplicity, we consider two-class problems here.

diagnosis tasks under the data scarcity scenarios. We also applied CF-aug to medical imaging for breast cancer detection to demonstrate the potential of CF-aug in other domains.

II. Background & Related Work

A. Speech-based Depression Detection

Recent work has explored speech-based deep-learning approaches to depression detection. Architectures such as fully-connected networks [7], [13], [14], RNNs [15], [16], CNNs [17], and Transformers [18] have been employed to capture latent representation of depression from speech. However, data scarcity poses a major challenge to these models, leading to bias and overfitting. Researchers have addressed this issue through techniques like data augmentation [9], sampling [19], [20], and transfer learning [17], [21]. Our work focuses on addressing data scarcity using data augmentation to enhance performance. A detailed comparison with other methods will be discussed in Section V-B.

B. Data Augmentation

Data augmentation is a common technique for improving model performance in machine learning tasks. As mentioned in Section I, data augmentation methods can be divided into two categories, depending on whether they generate samples from the same class or different classes.

1) Intra-class Augmentation: This type of augmentation generates new samples by applying transformations to existing samples within the same class. The goal is to create within-class variations to enhance the model's robustness and generalization abilities. In computer vision, traditional techniques in this category include image rotation, scaling, and flipping. For speech, changing pitch, adjusting the speaking rate, and adding noise are common augmentation techniques. For text data, it is a common practice to perform synonym replacement and random insertion. Additionally, generative models like Variational Autoencoders (VAEs) [22] and Generative Adversarial Networks (GANs) [23] are effectively for intra-class augmentation [8], [24], [25].

2) Inter-class Augmentation: This type of augmentation considers the relationship between multiple classes when generating new samples for all classes. Therefore, for each original sample, the variations in its augmented samples are not limited to one class but depend on its relationship with the samples in other classes. Our method falls into this kind.

Some inter-class data augmentation methods use generative adversarial networks, such as CycleGAN [26], to transform representations of one class to another [27]–[29]. For example, CycleGAN has been used to generate samples of minority classes to address the class imbalance problem by transferring samples from the majority class (e.g., happy) to the minority class (e.g., sad) [27]. However, adversarial models need careful tuning, and thus, it might not be optimal for scenarios with limited data. Another approach called Mixup [30] explores inter-class relationships by positing that linear combinations of input feature vectors should correspond to similar combinations of their associated output labels. However, this assumption may not hold in medical applications. This is because unlike computer vision, diseases often present with varying degree of severity, causing the mixed label to be noisy and misleading. For example, mixing a severe case with a healthy sample might yield a mixed label indicating a mild condition.

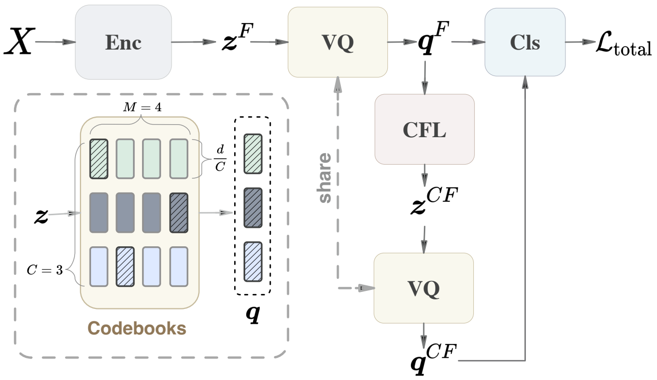


Fig. 1. Framework of CF-aug. Elements inside the dashed grey block illustrate the process of group-wise vector quantization.

C. Counterfactual Generation

Counterfactual generation refers to feature perturbation original feature that results in the model giving a different decision [11], [31]. Such generated counterfactual examples are used to explain the behavior of machine learning models, thereby helping model debugging, increasing models' interpretability, and enhancing human decision. Generally, given a machine learning model f and input feature vector $\boldsymbol{x}^F$, counterfactual feature vector $\boldsymbol{x}^{CF}$ can be obtained through optimizing the following objective [31]:

$$\boldsymbol{x}^{CF} = \arg\min_{\boldsymbol{x}} \mathcal{L}_{\text{cls}}(f(\boldsymbol{x}), y^{CF}) + \left|\left|\boldsymbol{x}^F - \boldsymbol{x}\right|\right|, \qquad (1)$$

where $\mathcal{L}_{\text{cls}}$ is a classification loss that encourages the counterfactual $\boldsymbol{x}^{CF}$ towards a different label y^{CF} other than $\boldsymbol{x}^F$'s ground truth y^F, and the second term keeps the counterfactual close to the original feature vector. Here, letters with superscripts F and CF denote original (factual) and counterfactual data, respectively. In the binary case, y_i^{CF} represents the counterfactual label of the i-th factual sample i, i.e., $y_i^{CF} = 1 - y_i^F$, where $y_i^F \in \{0, 1\}$. In this work, we leverage the same intuition as that in [11], [31] to change the original features to their counterfactual ones that lead to different outcomes for data augmentation, aiming to solve the data scarcity problem.

III. Methodology

A. Framework of CF-aug

Denote $\mathcal{D} = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^N$ as a dataset containing N sample-label pairs, where $\boldsymbol{X}_i$ and y_i are the feature vectors and the label of sample i, respectively. We aim to generate counterfactual features in the latent space by transforming the original data representation to their opposite class. The framework of CF-aug is shown in Fig. 1. The encoder Enc extracts feature representation $\boldsymbol{z}^F$ from the input feature vectors $\boldsymbol{X}$.[2] A group-wise vector quantization (VQ) module quantizes the latent vector $\boldsymbol{z}$ to discretized vector $\boldsymbol{q}$. Importantly, the counterfactual layer CFL changes the factual discretized vector $\boldsymbol{q}^F$ to the counterfactual vector $\boldsymbol{z}^{CF}$. The classifier Cls is used to diagnose disease based on the factual vector $\boldsymbol{q}^F$ and generated counterfactual vector $\boldsymbol{q}^{CF}$.

B. Group-wise Vector Quantization

We use a group-wise vector quantization (VQ) module to obtain a quantized vector $\boldsymbol{q} \in \mathbb{R}^d$ [32]. As shown in the dashed grey block in Fig. 1, the VQ module maps a continuous vector $\boldsymbol{z}$ to a discrete vector $\boldsymbol{q}$, composed of several entries sampled from the codebooks.

Given C groups (or codebooks), where the c-th group contains M entries $\boldsymbol{E}_c \in \mathbb{R}^{\frac{d}{C} \times M}$, the probability of selecting an entry from a group is determined by a Gumbel-Softmax distribution. Specifically, for the c-th group, the probability of selecting the j-th entry $\boldsymbol{e}_j \in \mathbb{R}^{\frac{d}{C}}$ from $\boldsymbol{E}_c$ given $\boldsymbol{z}^F$:[3]

$$p(\boldsymbol{e}_j|\boldsymbol{z}^F) = \frac{\exp((\log(\pi_j) + g_j)/\tau)}{\sum_{m=1}^{M} \exp((\log(\pi_m) + g_m)/\tau)}, \tag{2}$$

where π_j is the logit associated with the j-th entry $\boldsymbol{e}_j$, g_j is a Gumbel noise term that introduces randomness, and τ is a temperature parameter that controls the smoothness of the posterior distribution of $\boldsymbol{e}_j$. One entry is sampled from each codebook based on these probabilities, and the selected entries are concatenated to get $\boldsymbol{q} = \mathrm{concat}(\mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{C})$.

Why VQ? VQ can be effective in cases of data scarcity. This is because VQ acts as a form of prototype learning, where prototypes simplify the data by representing similar inputs with discrete entries. Using these entries makes the model less sensitive to noise and variations in the dataset, reducing the risk of overfitting and enhancing its robustness to distribution shifts in the test set. Notably, $\boldsymbol{z}^{CF}$ shares the same VQ module as $\boldsymbol{z}^F$, ensuring that both $\boldsymbol{q}^F$ and $\boldsymbol{q}^{CF}$ are generated from the same codebooks. This design explicitly promotes shared generation components between $\boldsymbol{q}^F$ and $\boldsymbol{q}^{CF}$, leading to unified representations. Consequently, VQ makes it easier to explore how different entries sampled from the codebooks affect the outcomes and prediction. Furthermore, this structure enables the generation of diverse outputs by allowing various combinations of sampled entries, thereby improving the model's capacity to produce a wide range of variations.

C. Model Training

1) Counterfactual Features Generation: Similar to [31], we obtain counterfactual features via optimizing the following loss:

$$\begin{aligned}\mathcal{L}_{cls} &= \mathcal{L}_{CE}^{F} + \mathcal{L}_{CE}^{CF} \\ &= \mathcal{L}_{CE}(\mathrm{Cls}(\boldsymbol{q}^F), y^F) + \mathcal{L}_{CE}(\mathrm{Cls}(\boldsymbol{q}^{CF}), y^{CF}),\end{aligned} \tag{3}$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, and class weights are used to address the class imbalance problem. Notably, the class weights of $\mathcal{L}_{CE}^{F}$ and $\mathcal{L}_{CE}^{CF}$ are swapped between classes, meaning $w_0^F = w_1^{CF}$ and $w_1^F = w_0^{CF}$, where w_0 and w_1 denote the weights for the negative and positive classes, respectively.

2) Auxiliary Constraint: To prevent the model from prioritizing the optimization $\mathcal{L}_{\text{CE}}^{F}$ over $\mathcal{L}_{\text{CE}}^{CF}$ through overfitting to the training data to reach a low classification loss, we introduce an auxiliary constraint $\mathcal{L}_{\text{aux}}$ to enforce similarity between $\mathcal{L}_{\text{CE}}^{F}$ and $\mathcal{L}_{\text{CE}}^{CF}$:

$$\mathcal{L}_{\text{aux}} = \max\left(\left|\mathcal{L}_{\text{CE}}^{F} - \mathcal{L}_{\text{CE}}^{CF}\right|, m\right), \tag{4}$$

where m is a margin that defines the maximum acceptable difference between $\mathcal{L}_{\text{CE}}^{F}$ and $\mathcal{L}_{\text{CE}}^{CF}$. Notably, given the same classifier Cls, $\mathcal{L}_{\text{CE}}^{F} \approx \mathcal{L}_{\text{CE}}^{CF}$ indicate that the overall feature distributions of the factual data $\boldsymbol{q}^F$ and generated counterfactual data $\boldsymbol{q}^{CF}$ are not significantly different. This is because the expected log-probability is a summary statistic of the feature distribution with respect to the classifier's predictions. Specifically, if

$$\mathbb{E}_{(y,\boldsymbol{q})\sim p^F} \log h(y|\boldsymbol{q}) \approx \mathbb{E}_{(y,\boldsymbol{q})\sim p^{CF}} \log h(y|\boldsymbol{q}), \tag{5}$$

where p_F and p_{CF} denote the joint distributions of the factual and counterfactual data and labels, respectively, then it implies that the overall feature distributions of the factual data $\boldsymbol{q}^F$ and counterfactual data $\boldsymbol{q}^{CF}$ are similar in how they influence the classifier's predictions. Here, $h(y|\boldsymbol{q})$ is estimated by the classifier Cls according to variational approximation theory, and it represents the probability of the true label y given the feature representation $\boldsymbol{q}$.

Overall, the total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{aux}}. \tag{6}$$

IV. Experimental Setup

A. Dataset

We evaluated CF-aug on two speech-based depression datasets. We also demonstrated CF-aug's potential in medical imaging using a breast cancer dataset.

a) DAIC-WOZ: The Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset [33] is widely used for studying and diagnosing depression. It comprises the clinical interviews of 189 participants who engaged with a virtual interviewer controlled remotely by a human interviewer. Participants were assessed using the PHQ-8 scale [34], with a score of 10 or above indicating depression. The dataset is officially divided into training, development, and test subsets. In this work, we validated our method on the development set and evaluated it on the test set.

b) MODMA: The Multimodal Open Dataset for Mental Disorder Analysis (MODMA) [35] is a depression dataset developed by Lanzhou University for research in mental disorder. MODMA comprises audio and EEG data collected from clinically diagnosed depressed patients and non-depressed controls. The depressed participants were recruited from the Second Hospital of Lanzhou University, while the non-depressed participants were recruited through public posters. The dataset includes the recordings of 52 subjects engaged in interviews,

[2] For speech, $\boldsymbol{X}$ can comprises a sequence of MFCCs, filterbank vectors, or frame-based feature vectors from a pre-trained model.

[3] For clarity, we omit the group index c and feature vector index i.

readings, and picture descriptions. Since there is no official dataset division, we adopted the test set division outlined in [21], and reserved 20% of the training data for validation.

c) BreastMNIST: The MedMNIST benchmark [36] offers a range of medical imaging datasets designed for rapid evaluation of machine learning methods in healthcare. Among its subsets, BreastMNIST focuses on breast cancer diagnosis and consists of 780 breast ultrasound images (546 training images, 78 validation images, and 156 test images) with a resolution of 28 × 28 pixels. In this study, we resampled the BreastMNIST dataset to include only 50 training samples in order to explore the versatility of CF-aug under data scarcity beyond its primary application in speech-based depression detection.

B. Data Preprocessing

For the DAIC-WOZ dataset, we extracted wav2vec features using a pre-trained model [37] and preprocessed the features as stated in [38]. For the MODMA dataset, we cut the waveform files to 3.84-second segments and extracted frame-based 80-dim filterbank features. For BreastMNIST, all images were preprocessed by subtracting the mean and dividing by the standard deviation to mitigate the impact of outliers.

C. Model Structures

This section introduces the model's structures and their ablated structures. We set $\tau = 1$ in Eq. 2, $M = 32$, and $C = 8$ in Fig. 1 for all models with a VQ module.

a) Depression model: We used the depression model to perform depression detection on DAIC-WOZ and MODMA. Enc, CFL, and Cls in the depression model comprise fully connected (FC) layers with the `tanh` activation function in their hidden nodes. There is a statistics pooling layer that concatenates the mean across time, the standard deviation across time, and the mean first-order difference between the successive feature frames of the output frame-based representations from the last layer of Enc to get segment-level vectors $\boldsymbol{z}^F \in \mathbb{R}^d$. Detailed model settings are shown in Table I.

b) Breast model: The breast model was trained to perform breast cancer classification on BreastMNIST. The Enc of our breast model consists of five convolutional (Conv) layers. Each Conv layer uses a 3×3 kernel with stride 1, followed by 2D batch normalization and `relu` activation. Max pooling (2×2 kernel, stride 2) was applied after the 2-nd and 5-th Conv layers. Following the Enc, we applied FC layers to reduce the dimensions of the latent vectors. The first FC layer transforms the 1024-dimensional vectors from the last Conv layer to 32 dimensions, followed by `relu` activation. The second FC layer transforms the 32-dimensional vectors to 64 dimensions, resulting in our final latent representation $\boldsymbol{z}^F$, i.e., $d = 64$. The details of the settings of Conv layers can be found in Table II.

c) Baseline (BL) model: We trained BL models using $\mathcal{L}_{CE}^F$ without augmenting counterfactual features, meaning only Enc, VQ and Cls in Fig. 1 were kept. Note that comparing BL to CF-aug is particularly valuable for assessing the impact of counterfactual sample augmentation, as both models use the same structure for inference.

TABLE I
THE NETWORK STRUCTURE (INPUT, HIDDENS, OUTPUT) OF DIFFERENT MODULES. THE NUMBERS INSIDE THE SQUARE BRACKETS INDICATE THE NUMBER OF HIDDEN NODES IN MULTIPLE LAYERS.

Module	**Network Structure**	
	Depression Model	**Breast Model**
Enc	512, 264, 16	-
Cls	48, [256, 64, 32], 2	64, 128, 2
CFL	48, [256, 128, 256], 48	64, 32, 64

TABLE II
THE SETTINGS OF CONV LAYERS IN OUR BREAST MODEL.

Index	**Layer**	**Kernel size**	**Stride**	**(Input, Output)**
1	Conv2d	3×3	1×1	(1, 16)
2	Conv2d	3×3	1×1	(16, 16)
3	Conv2d	3×3	1×1	(16, 64)
4	Conv2d	3×3	1×1	(64, 64)
5	Conv2d	3×3	1×1	(64, 64)

d) Regular baseline (R-BL) model: Since VQ is not commonly used in depression detection, we also reconfigured CF-aug into R-BL, consisting only Enc and Cls, which aligns more closely with typical deep learning-based depression detectors [7], [9], [19].

D. Metrics

We used a set of evaluation metrics, including Macro-averaged F1-score (MF1), F1-score for the positive class (F1-pos), F1-score for the negative class (F1-neg), and accuracy (Acc). For the depression detection task, we applied a majority voting strategy for the final classification. We conducted five independent runs of each experiment and reported the mean performance.

E. Network Optimization

We used the Adam optimizer for all experiments, training each model for 100 iterations. For the depression detection task, during each iteration, we trained the model on a batch of randomly sampled segments, with one segment selected from each speaker in the training set. We used a learning rate of 10^{-2} for the DAIC-WOZ dataset and 3×10^{-3} for the MODMA dataset. The batch size in the breast cancer classification task was 64, and the learning rate was fixed at 10^{-4}. We implemented a cosine scheduling strategy for learning rate adjustments, i.e., the learning rate was linearly increased from 0 in the first 10% of iterations, followed by a cosine-shape reduction in the remaining iterations. Early stopping was applied based on the MF1 on the validation set.

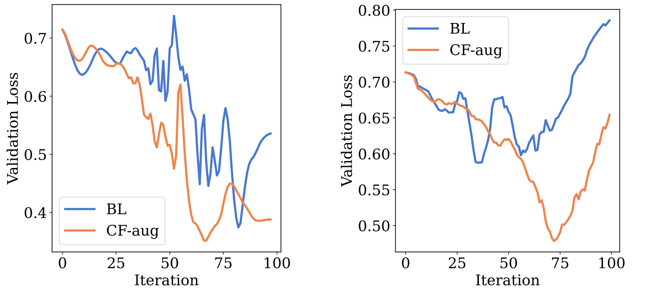
V. RESULTS

A. Effect of Counterfactual Features

The results for CF-aug and its ablated structures on DAIC-WOZ, MODMA, and BreastMNIST are shown in Tables III, IV, and V, respectively. Overall, CF-aug consistently outperforms its ablated structures, R-BL and BL, across all three datasets, including both the speech-based depression task and the breast cancer classification task. These results demonstrate

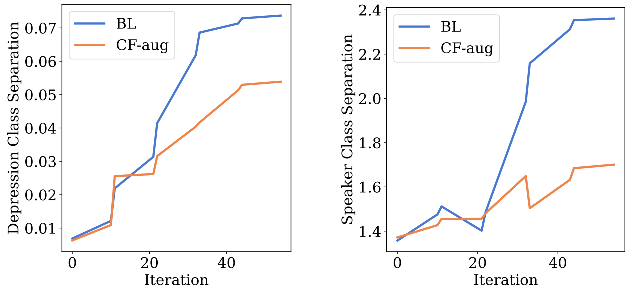
TABLE III
ABLATION STUDY ON DAIC-WOZ DATASET.

	Validation				**Test**			
	F1-pos	**F1-neg**	**MF1**	**Acc**	**F1-pos**	**F1-neg**	**MF1**	**Acc**
R-BL	0.513±0.110	0.771±0.038	0.642±0.055	0.693±0.034	0.374±0.104	0.730±0.051	0.552±0.052	0.628±0.051
BL	0.567±0.155	0.837±0.030	0.702±0.092	0.765±0.054	0.351±0.066	0.730±0.075	0.541±0.063	0.622±0.080
CF-aug	0.681±0.017	0.846± 0.014	0.764±0.013	0.793± 0.014	0.512±0.061	0.748±0.080	0.630±0.068	0.670±0.081



(a) MODMA (b) BreastMNIST

Fig. 2. Validation classification losses $\mathcal{L}_{ce}$ of BL and CF-aug for depression detection and breast cancer classification. The number of training samples used in Fig. 2(b) is 50.



(a) Depression Class Separation (b) Speaker Class Separation

Fig. 3. Plots of depression class separation and speaker class separation on Daic-woz dataset.

TABLE IV
ABLATION STUDY ON THE MODMA DATASET.

Method	**F1-pos**	**F1-neg**	**MF1**	**Acc**
R-BL	0.681±0.048	0.663±0.069	0.672±0.054	0.674±0.054
BL	0.725±0.028	0.790±0.045	0.758±0.035	0.763±0.038
CF-aug	0.770±0.043	0.810±0.036	0.790±0.038	0.793±0.037

TABLE V
ABLATION STUDY ON THE BREASTMNIST DATASET.

	F1-pos	**F1-neg**	**MF1**	**Acc**
R-BL	0.786±0.020	0.503±0.026	0.645±0.020	0.701±0.022
BL	0.796±0.023	0.506±0.032	0.651±0.025	0.712±0.027
CF-aug	0.815±0.018	0.511±0.024	0.663±0.019	0.732±0.018

the effectiveness of using augmented counterfactual features to improve model performance.

Reduce Overfitting. As shown in Fig. 2, CF-aug achieves lower validation losses on both tasks, suggesting that it can effectively mitigate the overfitting problem under data scarcity. This reduction in overfitting is likely due to the generation of counterfactual features, which introduces greater variability in the training data. This added variability acts as a regularization mechanism, preventing the model from memorizing specific patterns in the original dataset.

Reduce Bias. The work in [38] highlights the bias problem towards speaker features on DAIC-WOZ under data scarcity. We investigated the speaker class separability and depression class separability of BL and CF-aug in the latent space $\boldsymbol{q}^F$ to assess the impact of speaker features on both models. Class separability refers to how well the encoder distinguishes between samples from the same class and those from different classes [39]. Higher speaker class separability indicates a greater reliance on unintended speaker features, whereas higher depression class separability reflects a better capability to differentiate between depressed and non-depressed samples. As shown in Fig. 3, BL shows a consistent increase in speaker class separability, suggesting an unintended bias towards speaker-specific features. This bias indicates that BL may be overly focused on distinguishing between speakers rather than the target task of depression detection. Conversely, CF-aug maintains low speaker class separability, effectively mitigating speaker-related biases. Although BL exhibits higher depression class separability than CF-aug, we attribute this to an over-reliance on speaker features, which may lead to poor generalization across different speakers. CF-aug mitigates this bias by generating opposite-class samples in the latent space by recombining feature entries in the codebooks, effectively reducing the associations between certain features and the specific depression class. During this recombination process, CF-aug allows speaker features to be exchanged between different speakers, reducing the model's tendency to overfit to speaker-specific characteristics and ensuring a stronger focus on depression-related features. While recombining feature entries from codebooks has potentials to reduce bias by disrupting spurious correlations in the training set, this effect is not always guaranteed. Its effectiveness may vary depending on factors such as the nature of the task and the characteristics of the dataset. Consequently, controlling pre-identified confounding factors (or spurious features), such as speaker identity features in speech-based depression detection, based on prior knowledge, may yield more robust results.

B. Comparing with Other Methods

As shown in Tables VI and VII, CF-aug achieves state-of-the-art performance on both the DAIC-WOZ and MODMA datasets. FRAUG [9] augments samples by generating multiple views of the same speech segment through frame width and shift adjustments, helping the model focus on

TABLE VI
COMPARING WITH OTHER METHODS ON THE DAIC-WOZ DATASET.

Method	Validation			Test		
	F1-pos	F1-neg	MF1	F1-pos	F1-neg	MF1
FRAUG [9]	-	-	0.656	-	-	0.479
LECE [40]	0.692	0.818	0.755	-	-	0.553
SIDD [7]	0.741	0.866	0.805	0.481	0.721	0.601
Mixup [30]	0.675	0.796	0.736	0.466	0.680	0.573
CF-aug	0.681	0.846	0.764	0.512	0.748	0.630

TABLE VII
COMPARING WITH OTHER METHODS ON MODMA DATASET.

Method	F1-pos	Acc
DCL (Audio) [41]	0.615	0.615
GNN-SDA (Audio) [21]	0.766	0.788
Mixup [30]	0.753	0.770
CF-aug	0.770	0.793

invariant depression features across different views. However, this method is weak at introducing diversity and reducing bias, as it only creates variations from the same data without adding new samples. In the worst case, it can reinforce existing biases present in the original data and fails to generalize effectively. CF-aug outperforms FRAUG on DAIC-WOZ, likely due to introduced variation via combining entries selected from different feature groups and its potential ability to reduce bias.

Domain adaptation methods (DCL [41] and GNN-SDA [21]) are designed for transferring knowledge from a source domain to a target domain to enhance detection performance. CF-aug achieves performance comparable to DCL and GNN-SDA on the MODMA dataset while using less data. Unlike DCL and GNN-SDA, which typically require at least one additional dataset for adaptation, CF-aug does not rely on additional data, making CF-aug more efficient.

Given that speaker features are identified as a disturbing factor in the DAIC-WOZ dataset for depression detection [7], both SIDD [7] and LECE [40] aim to minimize the impact of these features on model decisions. As analyzed in Section V-A, CF-aug can also reduce the impact of speaker features. Importantly, it is not necessary for CF-aug to pre-identify these disturbing factors. CF-aug outperforms SIDD [7] and LECE [40], likely because it also reduces the impact of other unidentified factors and benefits from the increased diversity of data introduced by generated samples. However, directly controlling pre-identified disturbing factors, as done in SIDD [7] and LECE [40], remains a robust and straightforward approach.

We also implemented the Mixup method to compare with CF-aug, as both methods involve combining features for augmenting data. As mentioned in Section II-B, mixing different classes can introduce label noise, leading to misleading model training. We conducted experiments using Mixup on different classes and the same class. However, the models failed to learn meaningful patterns in all three datasets when mixing different classes, likely due to the introduction of label noise. Therefore, we only present the Mixup's results in Tables VI, VII, and VIII, focusing on cases where samples were mixed within the same class in the original space. The experimental results indicate that CF-aug consistently outperforms Mixup across all datasets. This performance gap can be attributed to two reasons: 1) Mixup's effectiveness will be limited if it only mixes samples from the same class, restricting its ability to establish smooth decision boundaries between different classes. This limitation could make it harder for the model to capture and handle uncertainty; and 2) Mixup mixes entire samples directly, which can limit its ability to generalize across diverse feature sets. In contrast, CF-aug can combine features from different groups more effectively, suggesting its flexibility in merging information, which contributes to its superior performance.

Moreover, we applied several standard augmentation techniques—such as Flip, Affine, and Mixup—to BreastMNIST and found that CF-aug consistently outperformed these techniques in terms of accuracy (see Table VIII). This demonstrates CF-aug's superior effectiveness and its potential for enhancing performance in medical imaging tasks.

TABLE VIII
COMPARING WITH OTHER AUGMENTATION METHODS ON BREASTMNIST.

	F1-pos	F1-neg	MF1	Acc
Affine	0.751±0.034	0.466±0.013	0.609±0.022	0.662±0.034
Flip (Horizontal)	0.771±0.009	0.526±0.021	0.645±0.014	0.690±0.012
Flip (Vertical)	0.781±0.015	0.509±0.026	0.645±0.016	0.697±0.016
Mixup	0.800±0.023	0.526±0.028	0.663±0.024	0.719±0.027
CF-aug	0.815±0.018	0.511±0.024	0.663±0.019	0.732±0.018

VI. DISCUSSIONS & CONCLUSIONS

Depression detectors often struggle to learn meaningful diagnostic features due to overfitting and bias caused by data scarcity. These incorrect predictions can have serious real-world consequences, highlighting the significance for developing robust models. In this study, we propose a CF-aug framework that generates counterfactual features for speech-based depression detection to address data scarcity. Experimental results show that CF-aug achieved state-of-the-art results on two depression datasets and effectively mitigated overfitting and bias problems. Additionally, we highlight CF-aug's potential in other medical diagnostics, including breast cancer detection in medical imaging.

We also discusses the limitations of CF-aug. Since the BreastMNIST dataset contains more training samples than depression datasets, it was used to evaluate the effectiveness of CF-aug with varying training set sizes.

a) Limited in generating out-of-distribution samples: CF-aug increases data diversity by generating samples by recombining entries in feature groups. This approach benefits datasets with limited natural variations, but it struggles to produce truly out-of-distribution (OOD) samples. Since larger datasets inherently possess greater natural variation, additional in-domain variations introduced by CF-aug will have a diminished impact. As a result, CF-aug's effectiveness is more pronounced in small datasets. As shown in Table IX, CF-aug is less effective with larger training sets (approximately 500 samples) compared to scenarios with fewer samples (50 and 200 samples). In addition, CF-aug might still underperform on OOD test sets. For example, CF-aug may struggle to detect depression in patients from countries outside the training set due to distributional shifts introduced by variations in language or accent.

TABLE IX

EFFECTIVENESS OF CF-AUG WITH VARYING TRAINING SET SIZES, AS MEASURED BY ACCURACY ON THE BREASTMNIST DATASET. "ALL" MEANS USING ALL TRAINING SAMPLES (546 IMAGES) PROVIDED IN BREASTMNIST.

	# of training samples		
	50	**200**	**All**
BL	0.712±0.027	0.755±0.025	0.841±0.013
CF-aug	0.732±0.018	0.785±0.009	0.849±0.022

b) Lack of causal knowledge integration: Similar to the limitation in [11], CF-aug generates counterfactual features without explicitly considering causal relationships in the data. The lack of causal knowledge can result in counterfactual features that may not be valid or interpretable in the real world. Addressing this issue will be left to future work.

c) Task and dataset-specific limitations of VQ: CF-aug relies on the VQ module for generation. However, the effectiveness of VQ can vary across tasks and datasets. For example, in fine-grained image segmentation tasks, such as detecting subtle variations in medical imaging for early disease diagnosis, VQ may lead to the loss of critical details essential for accurate analysis. Therefore, its application should be carefully evaluated based on the specific requirements of the tasks and datasets.

Although the VQ module may be less effective in some tasks and datasets, we want to highlight its potential in causal structure design. This is because VQ can be viewed as a form of soft intervention, contrasting with traditional interventions that involve setting a variable to a specific value [10]. It can serve as a tool for networks to perform interventions to infer potential causal relationships. Therefore, we recommend that future research in causal discovery consider incorporating VQ module(s) into their designs, as VQ may be valuable for identifying indicators and features that are robustly and causally linked to disease states from complex, high-dimensional data, such as speech and imaging.

[1] Jonathan Rottenberg. The prevalence of depression. *Depression*, 2021.

[2] Katie M Smith, Perry F. Renshaw, and John A. Bilello. The diagnosis of depression: Current and emerging methods. *Comprehensive Psychiatry*, 54 1:1–6, 2013.

[3] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[4] Klaus R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143, 1986.

[5] Gary Christopher and John MacDonald. The impact of clinical depression on working memory. *Cognitive Neuropsychiatry*, 10(5):379–399, 2005.

[6] Tomasz Rutowski, Amir Harati, Elizabeth Shriberg, Yang Lu, Piotr Chlebek, and Ricardo Oliveira. Toward corpus size requirements for training and evaluating depression risk models using spoken language. In *Proc. Interspeech*, pages 3343–3347, 2022.

[7] Lishi Zuo and Man-Wai Mak. Avoiding dominance of speaker features in speech-based depression detection. *Pattern Recognition Letters*, 173:50–56, 2023.

[8] Lu Yi and Man-Wai Mak. Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):172–184, 2022.

[9] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6267–6271, 2022.

[10] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[11] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc. Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[12] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[13] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 22(2):525–536, 2017.

[14] Xiaowei Zhang, Jian Shen, Zia ud Din, Jinyong Liu, Gang Wang, and Bin Hu. Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2265–2275, 2019.

[15] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Proc. Interspeech*, pages 1716–1720, 2018.

[16] Ying Shen, Huiyu Yang, and Lin Lin. Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251, 2022.

[17] Zhaocheng Huang, Julien Epps, Dale Joachim, Brian Stasak, James R. Williamson, and Thomas F. Quatieri. Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns. In *Proc. Interspeech*, pages 4561–4565, 2020.

[18] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. In *Proc. Interspeech*, pages 346–350, 2022.

[19] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proc. the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42, 2016.

[20] Andrew Bailey and Mark D Plumbley. Gender bias in depression detection using audio features. In *European Signal Processing Conference (EUSIPCO)*, pages 596–600, 2021.

[21] Tao Chen, Yanrong Guo, Shijie Hao, and Richang Hong. Semi-supervised domain adaptation for major depressive disorder detection. *IEEE Transactions on Multimedia*, 26:3567–3579, 2024.

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[24] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[25] Peiye Zhuang, Alexander G Schwing, and Oluwasanmi Koyejo. FMRI data augmentation via synthesis. In *IEEE International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1783–1787. IEEE, 2019.

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[27] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *Advances in Knowledge Discovery and Data Mining*, pages 349–360, 2018.

[28] Fang Bao, Michael Neumann, and Ngoc Thang Vu. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *Proc. Interspeech*, pages 2828–2832, 2019.

[29] Zhenghua Xu, Chang Qi, and Guizhi Xu. Semi-supervised attention-guided cyclegan for data augmentation on medical images. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 563–568. IEEE, 2019.

[30] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[31] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*, 2017.

[32] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

[33] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proc. the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, 2014.

[34] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, 2009.

[35] H Cai, Y Gao, S Sun, N Li, F Tian, H Xiao, J Li, Z Yang, X Li, Q Zhao, et al. MODMA dataset: A multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*, 2020.

[36] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2-A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[37] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech*, pages 3465–3469, 2019.

[38] Lishi Zuo, Man-Wai Mak, and Youzhi Tu. Promoting independence of depression and speaker features for speaker disentanglement in speech-based depression detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10191–10195, 2024.

[39] Lei Wang and Kap Luk Chan. Learning kernel parameters by using class separability measure. In *The Sixth Kernel Machines Workshop, in conjunction with Neural Information Processing Systems (NIPS)*, pages 1–8, 2002.

[40] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. *Computer speech & language*, 86:101605, 2024.

[41] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5351–5360, 2021.