Mitigating Confounding in Speech-Based Dementia Detection through Weight Masking

Zhecheng Sheng[†]   **Xiruo Ding**[*]   **Brian Hur**[*]   **Changye Li**[*]
Trevor Cohen[*]   **Serguei Pakhomov**[†]
[†]University of Minnesota   [*]University of Washington
[†]{sheng136, pakh0002}@umn.com

Abstract

Deep transformer models have been used to detect linguistic anomalies in patient transcripts for early Alzheimer's disease (AD) screening. While pre-trained neural language models (LMs) fine-tuned on AD transcripts perform well, little research has explored the effects of the gender of the speakers represented by these transcripts. This work addresses gender confounding in dementia detection and proposes two methods: the *Extended Confounding Filter* and the *Dual Filter*, which isolate and ablate weights associated with gender. We evaluate these methods on dementia datasets with first-person narratives from patients with cognitive impairment and healthy controls. Our results show transformer models tend to overfit to training data distributions. Disrupting gender-related weights results in a deconfounded dementia classifier, with the trade-off of slightly reduced dementia detection performance.

1 Introduction

Transformer-based models (Vaswani et al., 2017) have excelled in language and vision tasks, particularly bidirectional encoder models like BERT (Devlin et al., 2019) and its variants (Liu et al., 2019; Sanh et al., 2020; Lee et al., 2020; Qian et al., 2022), which enhance classification through rich text representations. As these models gain traction in clinical tasks like dementia detection, it is crucial to ensure fairness in their predictions, given the high stakes of clinical decision-making. However, most models are optimized for task-specific accuracy without accounting for biases present in fine-tuned datasets (Baldini et al., 2022; Bolukbasi et al., 2016; Hutchinson et al., 2020; Webster et al., 2021; de Vassimon Manela et al., 2021), leading to spurious correlations.

Efforts to mitigate these biases have focused on two main approaches. One involves task-agnostic methods that enforce fair representation learning (Kaneko and Bollegala, 2021; Cheng et al., 2021; Guo et al., 2022), while the other targets bias reduction in specific tasks using annotated data (Shen et al., 2021; Ravfogel et al., 2022; Gira et al., 2022; Zhu et al., 2023). A particularly challenging form of bias is confounding bias (Landeiro and Culotta, 2018), which arises when extraneous factors distort the relationship between the input language and the diagnostic outcomes. In spoken language-based dementia assessment, the existence of confounders can influence both linguistic patterns and disease prevalence, leading models to learn unintended associations. Despite growing awareness of bias in machine learning, confounding bias in *low-resource* domains like healthcare remains underexplored, where imbalanced datasets exacerbate the problem.

In this study, we investigate gender confounding bias in dementia detection using speech-based datasets. Prior research in Alzheimer's disease has identified sex as a prominent risk factor, with significant differences in dementia incidence observed between males and females (Beam et al., 2018; Podcasy and Epperson, 2016). While such pathological disparities are clinically meaningful, machine learning models that predict dementia from speech should make predictions independently of gender differences. A biased model may rely on gender-specific language cues rather than clinically relevant markers of cognitive decline, potentially leading to misdiagnoses and unequal performance across demographic groups.

To address this, we introduce two novel bias mitigation techniques inspired by the Confounding Filter (Wang et al., 2019): Extended Confounding Filter (ECF) and Dual Filter (DF). We evaluate these methods on two dementia speech datasets[1] widely

---

[1]Due to limited public datasets for Alzheimer's disease (AD) classification, we also considered ADReSS (Luz et al., 2020a) but excluded it due to its small size compared to the other two.

used in cognitive linguistic research (Li et al., 2022; Farzana and Parde, 2023). Our main contributions in this paper are as follows [2]:

1. We identified under-explored gender confounding bias in speech datasets for dementia.

2. We extended the Confounding Filter method to the Transformers architecture and demonstrated improvements in downstream task performance.

3. We introduced the Dual Filter as a simple yet effective weight masking algorithm that identifies and ablates parameters associated with the confounding bias in the entire model's network (vs. individual layers).

4. We showed that both proposed methods effectively reduce the False Positive Rate (FPR) and Statistical Parity (SP) gap between genders while maintaining relatively strong model performance under various distribution shifts.

2 Related Work

In recent years, transformer-based models have demonstrated promising performance in dementia detection using patient speech data (Hernandez-Dominguez et al., 2018; Cohen and Pakhomov, 2020; Luz et al., 2020b; Guo et al., 2021; Li et al., 2022). However, these models are susceptible to inductive bias due to the small size of publicly available datasets utilized in most studies. A key concern is that these models may learn gender-specific language patterns from male and female participants performing the same task, and subsequently use these differences to make dementia predictions, regardless of the participants' true cognitive status.

The methods we propose involves isolating and removing the influence of model weights associated with a confounding variable. As such, our work relates to prior efforts aimed at regularizing information encoded within transformer networks. One line of research explores weight isolation through disentangled learning (Zhang et al., 2021; Colombo et al., 2021), which require specialized loss functions to minimize information overlap between targets and sensitive attributes. Adapter sub-networks, on the other hand, regulate and control access to information from protected features (Hauzenberger et al., 2023; Masoudian

---

et al., 2024). Without introducing additional training objectives, Liu et al. (2024) propose a gradient integration method to identify neurons responsible for disparities in output logit distributions among demographic groups. Lee et al. (2019) and Sun et al. (2024) develop weight importance ranking algorithms to locate redundant weights for network pruning. These algorithms track neuron activations or loss outputs by masking certain weights within a layer and assigning importance scores based on a calibrated dataset.

While most of these methods focus on pretrained models—some exceptions serve as baselines in our experiments—our work specifically addresses bias learned during fine-tuning for a given task. Compared to prior approaches, our method for identifying influential weights integrates seamlessly into standard training, requiring no additional components or objective formulations, yet achieving significant bias mitigation across various distribution shifts in the data.

3 Methods

3.1 Confounding Filter

Deep learning models often recognize false signals from confounding factors, leading to sub-optimal performance in many real-world cases (Szegedy et al., 2013; Nguyen et al., 2015; Wang et al., 2017b,a). To address this issue, the Confounding Filter (Wang et al., 2019) was proposed to address confounding biases in models trained on electroencephalogram and medical imaging data. In this approach, a deep learning model is denoted as having two components: $g(\cdot; \theta)$, a representation learning network, and $f(\cdot; \phi)$, a classification network. The algorithm first optimizes the entire network by solving the following objective:

$$\hat{\theta}, \hat{\phi} = \underset{\theta, \phi}{\arg\min} \, \mathcal{L}(y, f(g(X); \theta); \phi),$$

where $\mathcal{L}$ denotes the loss function to be minimized.

In the second phase, assuming we have access to the confounder label m in the dataset, the algorithm localizes weights that are reactive to the confounding variable. This is achieved through tuning $f(\cdot; \phi)$ towards m while keeping $g(\cdot; \theta)$ fixed. During the second phase, updates in $\hat{\phi}$ are tracked and normalized after each batch. The sum of normalized updates is denoted as $\pi = \frac{1}{b} \sum_{i=1}^{b} |\Delta\phi_i|$ where b is the number of total batches in the second
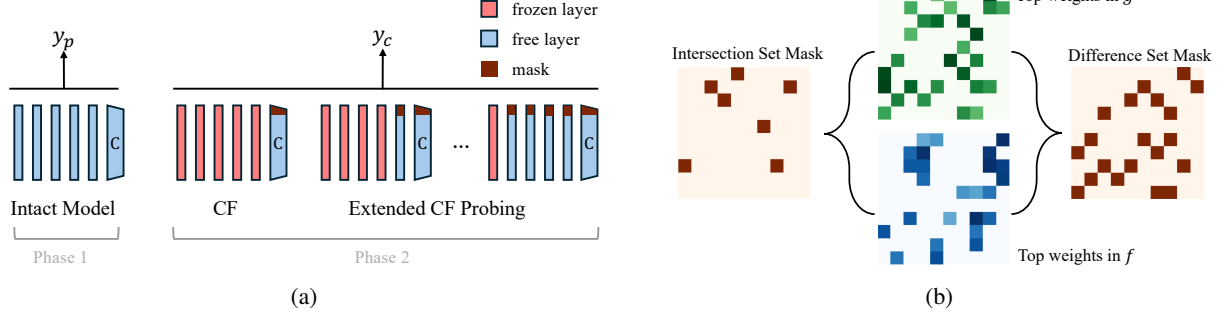
Figure 1: **(a)** Illustration of the Extended Confounding Filter (**ECF**) Probing framework for weights identification. **(b)** Illustration of the Dual Filter (**DF**) procedure to find weights to mask.

phase of training. The importance of each element in π is determined by their magnitude. A threshold function is then employed to get the mask:

$$M_i = \begin{cases} 0 & \text{if } \pi_i > \tau \\ 1 & \text{otherwise} \end{cases}$$

Here, τ is the kth percentile in π, where k is a hyperparameter. The element-wise product $\hat{\phi}' = \hat{\phi} \otimes M$ results in the confounder-mitigated network $f(g(X); \hat{\theta}); \hat{\phi}')$.

3.2 Extended Confounding Filter

While the original Confounding Filter algorithm has shown improvements over the baseline in some neural network architectures (Wang et al., 2019), its adaptation to transformer networks remains unexplored. Transformer-based language models (LMs) learn to generate distributional semantic representations (Vaswani et al., 2017) through the attention mechanism and positional encoding. By finetuning a pretrained LM, semantic information pertinent to a task of interest is dynamically stored across the transformer network layers.

Our hypothesis is that fixing $g(\cdot; \theta)$ when training for the confounder variable may not effectively capture the most confounder-associated weights within the transformer network. To test this hypothesis, we sequentially unfroze each layer in the transformer network, starting from the classification head down to the embedding layer and observed its impact on the prediction. This is different from the original Confounder Filter method, where only the classification head is trainable in the encoder model. We refer it as Extended Confounding Filter (ECF) in the paper.

The illustration of how ECF works is shown in Figure 2. Matrices $W_Q, W_K, W_V, W_O, W_1, W_2$ are tracked in a single transformer block, while W_{emb} and W_{cls} represent the token embedding matrix and classification weight matrix in a sequence

classification model, respectively. Similarly to the Confounding Filter, we start by training a classification model towards the primary label Y_p (Phase 1) and then continue training the model towards classifying the confounder label Y_c with layers sequentially unfreeze (Phase 2).
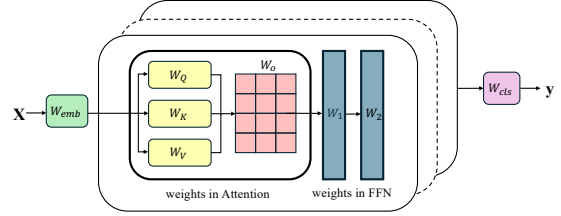

Figure 2: Tracked weights in the transformer network

By sequentially unfreezing different numbers of layers, we allow varying amounts of the model's parameter spaces to react to the information introduced during Phase 2 (Figure 1a). The sequential probing scheme follows the idea of the Confounding Filter but offers greater flexibility, as it allows partitioning of the classification network $f(x)$ and representation learning network $g(x)$ at different points. The change in model parameters $\Delta\phi_i$ is normalized within the matrix and recorded after each training batch. Following the Confounding Filter methodology, we restrict $\Delta\phi_i$ to each W in this probing procedure, and the threshold τ is calculated for each individual weight matrix. The probing step size is by layer. Masking matrices, derived from the threshold function, are applied to the tracked weight matrix from Phase 1 fine-tuning.

3.3 Dual Filter

We further relax the restriction in Phase 2 training of the ECF method, which mandates local masking and overlooks the dynamic interactions within the LM during fine-tuning. To address this, we propose Dual Filter, a method that tracks weight changes from two separate models initialized from the same checkpoint—one optimized for the primary out-

come (*dementia*) and the other for the confounder (*gender*). After obtaining change matrices π from both models, we utilize set operations to isolate weights that are most reactive to the confounder label during finetuning. Specifically, we choose the top $k\%$ most changed weights from the primary model f and the confounder model g, and take the intersection or the difference from these two weight sets to generate the mask matrices (Figure 1b). One could strategically apply the intersection set mask (M_I), the difference set mask (M_D), or the joint set ($M_I \cup M_D$) of both masks, which is equivalent to selecting the top k% most changed weights from the confounder model, depending on the dataset and task. This flexibility allows for precise bias mitigation, ensuring optimal trade-offs between fairness and performance tailored to specific applications. We formally describe the proposed method in Algorithm 1.

Algorithm 1 Dual Filter for weights masking

Input: pretrained LM: $f_0(x)$, $g_0(x)$; dataset: $\mathcal{D}(x, y_p, y_c)$; threshold: k
Output: Confounder-adjusted model $f(x; \theta')$
 1: Train $f_0(x; \theta) \mapsto y_p$, obtain weights change Δ_p and fine-tuned model $f(x; \hat{\theta})$.
 2: Train $g_0(x; \phi) \mapsto y_c$, obtain weights change Δ_c and fine-tuned model $g(x; \hat{\phi})$.
 3:
$$\Delta_{p,k} = \arg\max_{p \subseteq \Delta_p, |p|=k} \sum_{p_i \in \Delta_p} p_i$$
$$\Delta_{c,k} = \arg\max_{c \subseteq \Delta_c, |c|=k} \sum_{c_i \in \Delta_c} c_i$$
 4: $M_I \leftarrow \Delta_{p,k} \cap \Delta_{c,k}$, $M_D \leftarrow \Delta_{c,k} \setminus \Delta_{p,k}$
 5: Pick mask $M \in \{M_I, M_D, M_I \cup M_D\}$
 6:
$$\theta' \leftarrow \hat{\theta}_i = 0 \quad \forall i \in M$$

3.4 Other Baselines

We include two recent baseline adapter models, CONGATER (Masoudian et al., 2024) and MOD-DIFFY (Hauzenberger et al., 2023) using their publicly available code[3,4], both of which address biases learned during fine-tuning. These methods append additional bias-mitigate modules into a network and update their weights through a joint loss function. We run these two methods using their

---

[3] https://github.com/ShawMask/DebiasingConGater
[4] https://github.com/CPJKU/ModularizedDebiasing

---

default configurations with a range of hyperparameters (See Table S3) to compare their fairness-performance trade-offs with the proposed methods.

4 Evaluations

Confounding Shift One fundamental assumption in machine learning is that the test and training datasets are from the same distribution. However this assumption is often violated in real world applications resulting in distribution shifts. One specific form of distribution shift is sub-population shift (Cao et al., 2019; Cai et al., 2021), where the training distribution differs from the deployment distribution. A model trained on such shifted data tends to learn spurious correlations with the majority class, resulting in poor performance when applied to data with a class distribution different from that of the training set (Yang et al., 2023).

While the sub-population shifts are determined by the product of group attributes and the label, and the group attributes are not independent of the label, it is a special type of dataset shift referred to as *Confounding Shift* (Landeiro and Culotta, 2018). Formally, confounding shift exists when two conditions are met: (i) a confounding variable Y_c exists that impacts both X and Y_p through distributions $P(X|Y_c)$ and $P(Y_p|Y_c)$ through the backdoor path in a causal graph (Pearl, 2009); (ii) a subpopulation distribution $P_{train}(Y_p|Y_c)$ is different from $P_{test}(Y_p|Y_c)$ (Landeiro and Culotta, 2018).

To quantitatively assess the degree of confounding shift, we use a framework proposed by Ding et al. (2024) in our experiments. This allows us to perturb the target variable and confounding variable distributions in both training and test splits to different degrees through sampling from the original dataset. Under this framework, we consider a dataset with a binary target and binary confounder, the joint distribution $P(Y_p, Y_c)$ governed by the following quantity: $P(Y_c = 1), P(Y_p = 1), P(Y_p = 1|Y_c = 1), P(Y_p = 1|Y_c = 0)$. Next Ding et al. (2024) introduced a positive auxiliary variable $\alpha = \frac{P(Y_p=1|Y_c=1)}{P(Y_p=1|Y_c=0)}$, which serves as a knob for controlling the degree of subpopulation shift. By setting different α values, we control the source of the positive examples. If we hold $P(Y_c = 1)$ and $P(Y_p = 1)$ constant, we can vary α_{train} and α_{test} to create a mixture of datasets with various degrees of shift for model evaluation. Details are described in Section 5.2.

Fairness Fairness in machine learning seeks to ensure that models make unbiased decisions and perform equally well across different demographic groups. One widely used notion of group fairness is statistical parity, which emphasizes equal outcomes at the population level (Dwork et al., 2011). In binary classification tasks with a binary group attribute G and binary outcome Y, statistical parity is measured by the absolute difference or ratio between $P(\hat{Y} = 1 \mid G = 1)$ and $P(\hat{Y} = 1 \mid G = 0)$. A smaller disparity indicates more equal treatment across groups in the model's predictions.

Beyond statistical parity, other fairness metrics incorporate ground-truth labels to assess prediction accuracy across groups. For example, Equal Opportunity compares true positive rates between subgroups to evaluate whether the model performs equally well for individuals who belong to the positive class (Hardt et al., 2016; Romano et al., 2020).

In our study setup, test set distributions can vary with the parameter α, which governs the prevalence of dementia. This variation affects the base rate of dementia. To address this, we focus on the false positive rate (FPR), defined as $P(\hat{dementia} = 1 \mid \text{gender}, \text{dementia} = 0)$—the model's predicted dementia probability among healthy individuals. We measure the absolute difference in FPR between gender groups, denoted as ΔFPR, to assess fairness in terms of error rates.

Additionally, to evaluate whether our method reduces statistical parity, we examine the absolute output probability difference $|P(\hat{dementia} = 1 \mid \text{gender=F}) - P(\hat{dementia} = 1 \mid \text{gender=M})|$, denoted as ΔSP, on a balanced test set. This allows us to determine whether the model produces equitable predictions for male and female participants when the underlying training distribution is controlled.

5 Experimental Setup

5.1 Dataset

DementiaBank (DB) The benchmark dataset used for our experiments is the Pittsburgh Corpus from DementiaBank (Becker et al., 1994; MacWhinney, 2007) This corpus is a widely used resource in the fields of computational linguistics and dementia studies. It provides detailed speech and language data from elderly participants with dementia as well as healthy controls. Notably, the Pittsburgh Corpus includes responses to the *Cookie Theft* picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Ka-

plan, 1983). The dataset comprises 548 examples collected from longitudinal records of 290 participants. To ensure the transcripts accurately reflect the diagnosis label, we selected the last transcript for each patient as input for our model (183 female vs 107 male).

Carolinas Conversation Collection (CCC) The Carolina Conversations Collection (CCC) (Pope and Davis, 2011) differs from DB by sourcing English conversational interviews rather than neuropsychological tasks. The corpus contains 646 interviews from 48 cognitively normal elderly individuals and 284 with dementia, with participants potentially having multiple interviews. These conversations focus on health-related autobiographical narratives and have been widely used in psycholinguistic NLP research (Nasreen et al., 2021; Li et al., 2022; Farzana and Parde, 2023). Our study uses 394 transcripts from 70 interviewees with available gender information (323 female vs 71 male).

5.2 Experiments

We start by examining whether a text classification model will recognize gender confounding bias from speech data. We fine-tuned a BERT-base model (Devlin et al., 2019) on the complete dataset and assessed its performance across gender-specific subgroups. We ran the experiments using 5-fold cross validation with 3 repeats on both the original dataset, and a perfectly balanced dataset created by down-sampling the more prevalent category. Mann-Whitney-Wilcoxon test were performed between male and female and the result is shown in Table 1 - performance discrepancies were observed among male and female examples across multiple runs. This result shows that there exists confounding by gender in the dementia detection task which is independent of the gender distribution in the dataset. This suggests that the gender of the speaker influences the language they use to complete the Cookie Theft picture description task or the phone interviews, and confound the dementia signals during model fine-tuning. Hereby, we further investigate this confounding by gender effects in dementia detection and evaluate our proposed deconfounding methods.

Dataset Perturbation As described in Section 4 , we manipulated the conditional distribution of dementia by gender in our dataset through random sampling, creating a series of datasets with varying levels of confounding shift. In our experiments, de-
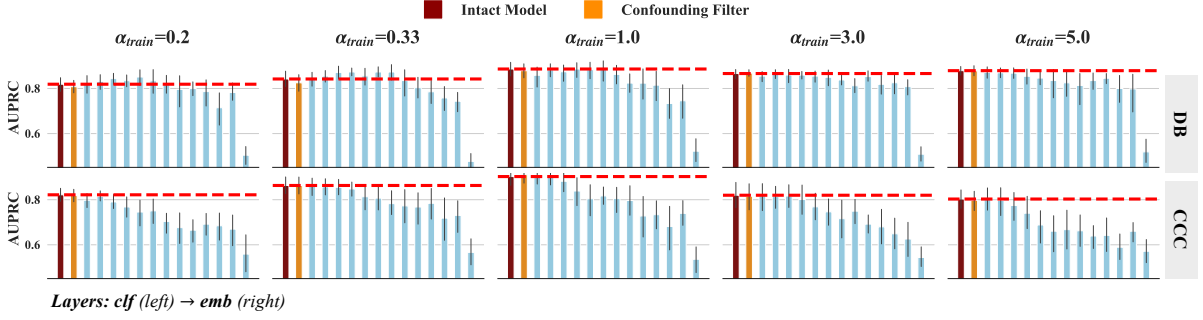
Figure 3: ECF filtering with 15% masking rate across different confounding shifts from two dataset.

dataset	setup	abs mean diff	p-value
DB	Original	0.055	< .001
	Balanced	0.068	< .001
CCC	Original	0.152	0.002
	Balanced	0.102	0.007

Table 1: Two sided Mann-Whitney-Wilcoxon test results of male and female dementia prediction performance (AUPRC) across different setups.

mentia cases and female participants are coded as 1, respectively. We fixed $P(\text{gender} = 1) = 0.5$ and $P(\text{dementia} = 1) = 0.5$ in both the training and test sets to ensure fair comparisons across different configurations. This way, the dataset is balanced with respect to both dementia and gender. Then we adjusted the value of $\alpha = \frac{P(\text{dementia}|\text{female})}{P(\text{dementia}|\text{male})}$ to create an imbalance in the source of dementia cases (sub-population shift). If $\alpha > 1$, more dementia cases are drawn from females, while $\alpha < 1$ indicates the opposite. The further α is from 1, the more severe the imbalance. To evaluate the model's robustness to confounding shifts, the model is trained on one α_{train} value and tested on its reciprocal value $\alpha_{\text{test}} = \frac{1}{\alpha_{\text{train}}}$, simulating an extreme shift in the test set compared to the distribution the model was exposed to during training. The selection of $\frac{1}{\alpha_{\text{train}}}$ is entirely arbitrary and is intended solely to illustrate the shift magnitude and establish the testbed for evaluation. Models are trained for 20 epochs on 480 training examples, validated on 120 examples and evaluated on 150 examples for each configuration. Among them, the training set and validation set are sampled from α, while the test set is sourced from $\frac{1}{\alpha}$. The best checkpoint is selected based on the AUPRC on the validation set, using early stopping to prevent overfitting.

Extended Confounding Filter The encoder model we used for dementia detection is BERT-base, with 12 encoder layers and 12 attention heads in each layer. Once we obtain the dementia fine-tuned model $f(x)$ after the first Phase, we

take a snapshot of the parameters and only make some parts of it trainable towards the gender label in the second Phase. The trainable layers start with cls, and one layer is sequentially added to the trainable set. Eventually, the trainable set becomes $\{cls, layer12, layer11, ..., layer1, emb\}$ and spans the whole network. Then for each trainable set, f_d is trained toward gender prediction. We ranked the weights that changed in each layer and selected the top 15% of the weights with the most significant changes in each layer to mask (Figure 1a), following the approach in Wang et al. (2019). Then we evaluated the masked models. We include results of choosing different masking ratios for ECF in Appendix C.2.

Dual Filter In the Dual Filter approach, we track the global weight change throughout the model's architecture. The classification head is exempt from tracking as it is training toward two different tasks and the weights in the classification head are assumed to have the most significant change compared to the rest of network. We first obtain two lists of weights change matrices from $f(x)$ and $g(x)$, using the same approach as Extended CF. Then we rank and select the top $k\%$ weights by their locations in the network. A sequence of k values are tested, ranging from 0 to 60 and step size of 1. Then three sets ($M_I, M_D, M_I \cap M_D$) are calculated and applied to $f(x)$ to create the masked model. Note when training toward gender in both Extended CF and Dual Filter, we select only *non-dementia* cases to let the model learn from texts that are representative of the gender differences. Consequently, only healthy cases are used in the evaluation.

6 Results

6.1 Extended Confounding Filter

Figure 3 presents the results of the Extended Confounding Filter (ECF). The red dotted line repre-

sents the performance of the intact model, while the bars illustrate models where weights are progressively eliminated, layer by layer, from left to right until the embedding layer is reached (the rightmost bar). The orange bar represents the original Confounding Filter approach, where only the classification head is trained in the second phase and then masked.

We observe that simply applying the Confounding Filter to the classification layer is insufficient to mitigate confounding bias. Propagating masks layer by layer helps maintain or even improve dementia classification performance. Specifically, the model remains resilient in dementia detection when gender-associated weights are removed from the upper layers, with no significant performance drops occurring until weights are ablated from lower layers. This aligns with prior work (Li et al., 2024), which reports similar robustness in linguistic feature encoding. Interestingly, in some cases, removing gender-associated weights from certain layers improves AUPRC compared to the intact model, suggesting potential patterns that warrant further investigation. Across both datasets, a "ladder" effect emerges due to confounding shift: models trained and tested on the same distribution ($\alpha = 1$) achieve the highest performance, while the performance declines as α_{train} deviates from 1.

In particular, the token embedding layer plays a critical role in dementia detection—our experiments show that removing even a small fraction of its weights drastically alters the model's performance. Further research is needed to better understand its contribution when using transformer models for dementia detection.
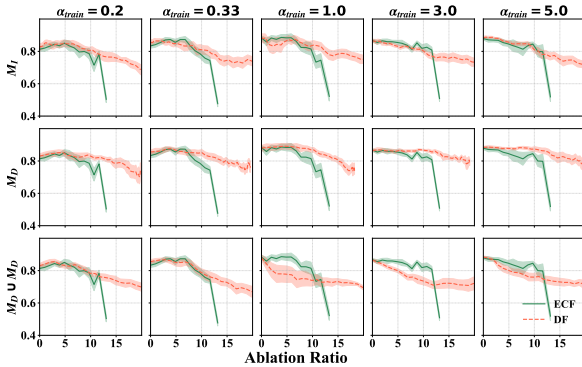
6.2 Dual Filter

Figure 4: AUPRC (y-axis) on ECF and DF for different α_{train} configurations. The x-axis represents the % of ablated weights through the whole network.

In Figure 4, we visualize the dementia prediction

performance change on the DB dataset as we apply three different types of mask to the original model and gradually increase the masking ratio. The results from ECF with 15% layer-specific masking ratio are added for comparison. The plot shows the relation between how many weight entries are ablated within the whole network against model AUPRC. The rows indicate three types of masks that are generated by Dual Filter and the columns indicate the specific α_{train} configurations that control the distribution shift.

Next, we show the absolute False Positive Rate difference (ΔFPR) between females and males calculated under both ECF and DF methods. Figure 5 shows the change in FPR gaps as the ablation ratios increase for all three types of masks. The mask type is indicated in rows while the columns represent different α_{train}. Similar trends are also observed in the CCC dataset, shown in Appendix C.1.
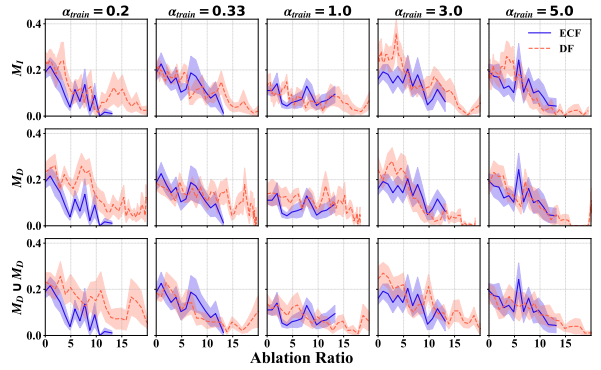

Figure 5: ΔFPR (y-axis) on ECF and DF for different α_{train} configurations.
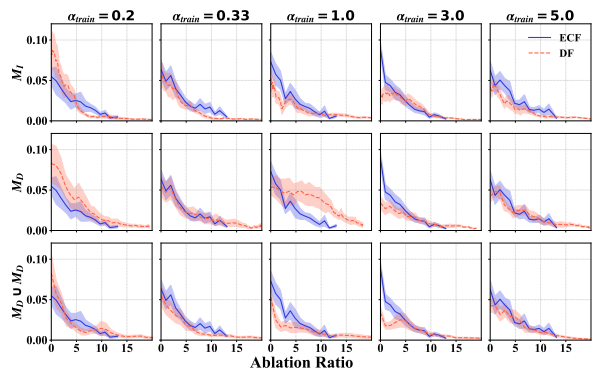

Figure 6: ΔSP (y-axis) on ECF and DF for different α_{train} configurations.

The results in Figure 4 show that all three masking strategies display similar trends under varying confounding shift configurations. Both the M_I and M_D masks exhibit stronger resilience across different ablation ratios compared to the ECF baseline. In contrast, the union mask ($M_I \cup M_D$), which
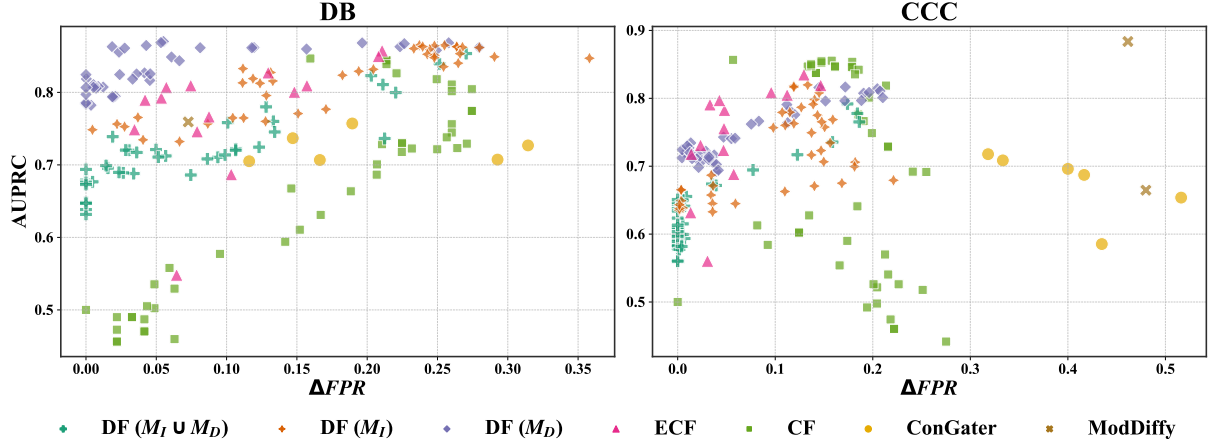
Figure 7: AUPRC vs. ΔFPR among different methods for two datasets.

removes a larger set of weights, exhibits comparable or slightly reduced resilience compared to ECF across different α settings. This suggests that removing all gender-related weights without considering their association with dementia labels impair model performance under distributional shifts.

Paired with Figure 5, we observe that in certain intervals of the ablation ratios, the performance remains stable while the fairness metric improves. For example, at $\alpha = 0.2$ in the DB dataset, removing 10% of the weights from M_I (the intersection mask) achieves model AUPRC at 0.80, which drop only slightly from the original model (0.82), while the FPR difference between male and female drops from 0.22 to 0.03. These results suggest an entanglement between weights responsible for dementia detection and those associated with gender, particularly in weight entries that undergo the most change across different layers of the network. We further analyze this entanglement across layers under different confounding shift settings in Appendix D.2. Figure 5 also suggests that FPR gaps are more severe in the original model under extreme confounding shifts.

In addition, we assess model fairness on the DB dataset using statistical parity, with the evaluation conducted on a balanced test set ($\alpha = 1$). As shown in Figure 6, both proposed weight masking methods substantially reduce the prediction disparity between male and female participants across different training data shifts. This improvement also suggests that the methods successfully identify and suppress gender-related weights in the model, leading to more equitable predictions.

6.3 Comparisons with other methods

We train different models on data with an arbitrarily selected setting with $\alpha_{train} = 3$ to represent a con-

founding shift case. That means the dementia cases for training are sourced three times more from the female cohort than the male cohort. We then test on samples from the configuration of $\alpha_{test} = \frac{1}{3}$. We evaluate the results on ECF, DF and other baselines using the AUPRC-ΔFPR curve (Figure 7), in which the upper-left points represent an ideal classifier, with both high accuracy and fairness.

Our experiments show that ECF achieves the best trade-off on the CCC dataset, while DF (M_D) outperforms the other methods on the DB dataset. Across both datasets, our methods consistently outperform adapter-based baselines and naive Confounding Filters, demonstrating their effectiveness in mitigating confounding shifts. Notably, compared to naive Confounding Filters, our methods achieve a more favorable trade-off, maintaining higher AUPRC at the several FPR disparity levels.

Additionally, we note that weight masking approaches offer a more fine-grained trade-off trajectory than loss-optimization methods, providing greater flexibility across various use cases.

7 Discussion

From the experiments, we conclude that both ECF and DF effectively mask gender-related weights within a BERT-base model, improving gender parity in outcomes while maintaining comparable performance in dementia detection under various degrees of confounding shifts. Comparisons with existing baselines demonstrate that our proposed methods achieve a more favorable trade-off between model performance and gender parity. We further discuss the broader impact of the work.

Clinical Implication While this work exclusively focuses on the gender-confounding issue in dementia, its broader implications extend to

other medical AI applications. Unbalanced data is common in many medical settings, and inductive bias in deep learning models can lead to misdiagnoses or uneven treatment recommendations during model inference. Our findings emphasize the need for proactive bias mitigation strategies in low-resource domains like dementia research, and future work should explore extending these deconfounding methods to additional clinical variables.

Generalizability Both the ECF and DF methods are model-agnostic and can be applied to any transformer-based architecture. Additionally, both methods can be adapted for non-binary confounders by formulating them as a multi-class classification task during model fine-tuning. Although this paper focuses on a single clinical task with a binary target and a binary confounder,the proposed framework can be extended to more complex confounding scenarios.

Scalability ECF suffers from scalability issues due to its retraining process at each layer to exploit the trade-offs, while DF is more computationally efficient, requiring only two fine-tuning steps. Therefore, its computational complexity scales linearly with the size of the dataset, making it more practical for larger datasets and models.

8 Conclusion

In this paper, we address gender confounding bias learned during model fine-tuning and propose two model-agnostic methods for filtering confounder-associated weights in transformer neural networks. We apply these methods to dementia detection tasks, demonstrating their potential utility in clinical practice. Our findings indicate that unaddressed confounding shifts can degrade model performance even when the overall label and group distributions are balanced. Experimental results compare the identification of gender-associated weights both layer-wise and across the entire model. Both methods show minimally degraded performance on the dementia detection task while reducing gender bias. We observed non-monotonic responses across layers, suggesting further investigation is needed to understand the inner workings of even small transformer models. Lastly, we note that ensuring fairness and maintaining model performance often involve trade-offs, and real-world decisions should consider multiple factors, including bias tolerance and use-case specifics.

Limitations

Dataset The experiments of our proposed methods are only conducted on two relatively small datasets; therefore, generalizability to other datasets will need to be further investigated. In addition, given the small data size, manifesting different levels of confounding shift requires repetitive sampling to meet the desired subgroup distribution. Thus the resultant dataset contains a significant amount of duplicates that may reduce the strength of the findings and would need to be replicated in larger datasets.

Methods In ECF methods, even though the approach we take is the most straightforward and allows the model to absorb unidirectional effects, we ignore the possibility of other combinations of layer freezing inside the network and leave it to be explored in future work.

Experiments While we acknowledge BERT-base as a good starting point for investigation, we did not include other encoder models in this work. Also, while we briefly discussed some other weight importance measurements to isolate weights that impact certain outputs, we did not include and compare them with our current approach for deconfounding mainly due to the publicly available code for other prior work not being model-agnostic.

Ethical Statement

Dataset Privacy The two datasets utilized in this study are publicly available (upon request) and have been fully de-identified. 1) DementiaBank transcripts: These are collected from picture description tasks in cognitive tests designed to assess structural language skills. 2) Carolina Conversations Collection transcripts: These are derived from interviews about life stories or personal well-being. While these datasets include demographic information about the subjects, they are considered a health-related speech dataset instead of medical dataset in general sense, and the privacy concerns associated with their use are minimal.

Potential Downstream Risk While the deployment of model-based dementia screening tools has the potential to support timely interventions and improve patient outcomes through early detection of cognitive decline, these benefits must be carefully weighed against ethical and practical concerns. For example, false positives of model predictions may

cause distress and anxiety to patients or lead to premature clinical decisions; Such models can also be applied for non-medical purposes such as monitoring cognitive status or unregulated cognitive assessments of vulnerable populations, which may aggravate discrimination against those groups.

Gender Clarification The datasets used in this study include only participants identified as male or female. In the DB dataset, gender labels are inferred from participants' reported biological sex, whereas in the CCC dataset, gender is explicitly provided in the metadata. We acknowledge that equating gender with biological sex in the DB dataset is a strong assumption, and this limitation warrants further investigation.

Acknowledgment

References

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. *Preprint*, arXiv:2108.01250.

Christopher R Beam, Cody Kaneshiro, Jung Yun Jang, Chandra A Reynolds, Nancy L Pedersen, and Margaret Gatz. 2018. Differences between women and men in incidence rates of dementia and alzheimer's disease. *Journal of Alzheimer's disease*, 64(4):1077–1083.

JT Becker, F Boller, OL Lopez, J Saxton, and KL McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Tianle Cai, Ruiqi Gao, Jason D. Lee, and Qi Lei. 2021. A theory of label propagation for subpopulation shift. *Preprint*, arXiv:2102.11203.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer's type. *Preprint*, arXiv:2005.03593.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. *Preprint*, arXiv:2101.09688.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiruo Ding, Zhecheng Sheng, Meliha Yetişgen, Serguei Pakhomov, and Trevor Cohen. 2024. Backdoor adjustment of confounding by provenance for robust text classification of multi-institutional clinical notes. In *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 923–932.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. *Preprint*, arXiv:1104.3913.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Harold Goodglass and Edith Kaplan. 1983. *Boston Diagnostic Aphasia Examination Booklet*. Lea & Febiger, Philadelphia.

Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, 3.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Preprint*, arXiv:1610.02413.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and on-demand bias mitigation with attribute-removal subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6192–6214, Toronto, Canada. Association for Computational Linguistics.

Luis Hernandez-Dominguez, Samuel Ratté, Basilio A. Sierra, and Jesus A. Roche-Berges. 2018. Computer-based evaluation of alzheimer's disease and mild cognitive impairment using lexical and syntactic information. *Journal of Alzheimer's Disease*, 63(2):709–719.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *Preprint*, arXiv:2005.00813.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *Preprint*, arXiv:2101.09523.

Virgile Landeiro and Aron Culotta. 2018. Robust text classification under confounding shift. *J. Artif. Int. Res.*, 63(1):391–419.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. Snip: Single-shot network pruning based on connection sensitivity. *Preprint*, arXiv:1810.02340.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Changye Li, Zhecheng Sheng, Trevor Cohen, and Serguei Pakhomov. 2024. Too big to fail: Larger language models are disproportionately resilient to induction of dementia-related linguistic anomalies. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6363–6377, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020a. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *Preprint*, arXiv:2004.06833.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020b. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China.

Brian MacWhinney. 2007. The talkbank project. In *Creating and Digitizing Language Corpora*, pages 163–180. Palgrave Macmillan, London.

Shahed Masoudian, Cornelia Volaucnik, Markus Schedl, and Navid Rekabsaz. 2024. Effective controllable bias mitigation for classification and retrieval using gate adapters. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2434–2453, St. Julian's, Malta. Association for Computational Linguistics.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, Singapore and Online. Association for Computational Linguistics.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Judea Pearl. 2009. *Causality*, 2nd edition. Cambridge University Press, Cambridge, UK.

Jessica L. Podcasy and C. Neill Epperson. 2016. Considering sex and gender in alzheimer disease and other dementias. *Dialogues in Clinical Neuroscience*, 18(4):437–446. PMID: 28179815.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *Preprint*, arXiv:2205.12586.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. *Preprint*, arXiv:2201.12091.

Yaniv Romano, Stephen Bates, and Emmanuel J. Candès. 2020. Achieving equalized odds by resampling sensitive attributes. *Preprint*, arXiv:2006.04292.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *Preprint*, arXiv:2109.10645.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. *Preprint*, arXiv:2306.11695.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Haohan Wang, Akshay Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017a. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *IEEE International Conference on Multimedia and Expo*.

Haohan Wang, Bhiksha Raj, and Eric P Xing. 2017b. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

Haohan Wang, Zhenglin Wu, and Eric P. Xing. 2019. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 24, pages 54–65.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. *Preprint*, arXiv:2010.06032.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *Preprint*, arXiv:2402.05162.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023. Change is hard: A closer look at subpopulation shift. *Preprint*, arXiv:2302.12254.

Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. 2021. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. 2023. Debiased fine-tuning for vision-language models by prompt regularization. *Preprint*, arXiv:2301.12429.

A Evidence of Confounding

A.1 Gender bias in dementia detection

(a) AUPRC results from DB   (b) AUPRC results from CCC

Figure S1: Performance discrepancy between male and female in two datasets using the BERT-base model.

A.2 Dataset Statistics

DB Dataset				CCC Dataset		
Gender	Label	Count		Gender	Label	Count
Female	0	57		Female	0	220
Female	1	126		Female	1	103
Male	0	41		Male	0	42
Male	1	66		Male	1	29

Table S1: DB and CCC Datasets gender-label counts

B Experiments Setup

B.1 Finetuning hyperparameters

	parameter	value
Data Size	Train	480
	Validation	120
	Test	150
Hyperparameters	Max sequence length	256
	Number of epochs	20
	Early stopping tolerance	5
	Early stopping metric	AUPRC
	Optimizer	AdamW
	Scheduler	Linear
	Warm up steps	50
	Learning rate	1e-5

Table S2: Model and Data Configurations for Finetuning

B.2 Hyperparameter selection for baseline methods

In Table S3, we present the selected hyperparameters for the two adapter-based baseline methods along with their corresponding definitions.

C Additional Results

C.1 CF vs ECF on CCC dataset

In this section we show the results for EF and DF comparisons for AUPRC and ΔFPR against the ablation ratio. Figure S2 demonstrates the ablation

effects on the AUPRC metric and Figure S3 shows the effect for the absolute FPR difference between female and male. The results suggest both methods work effectively on the CCC dataset.
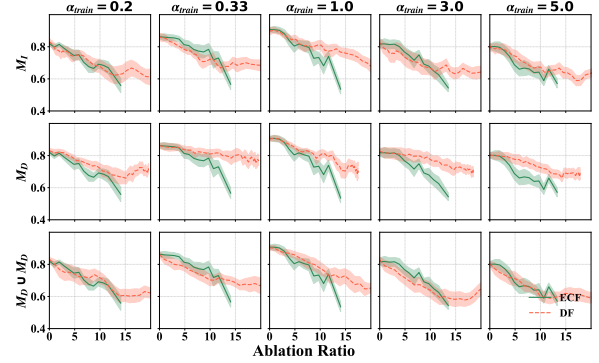

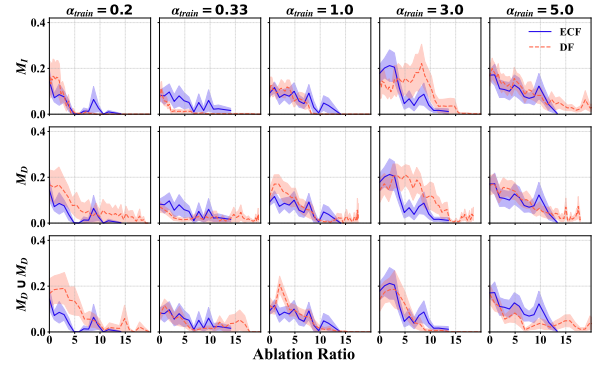Figure S2: Relationship between AUPRC and weights ablation on CCC dataset for ECF and DF.

Figure S3: Relationship between ΔFPR and weights ablation on CCC dataset for ECF and DF.

C.2 ECF with different masking ratios

In this section, we present ECF results with different masking ratios (5%, 25% and 35%). Figrue S4, S5, and S6 demonstrate the results. We can observe that as the masking ratio increases, the model performance on dementia detection regarding AUPRC drops significantly after several layers. We then assess whether masking only 5% of the weights in ECF can effectively reduce FPR gaps. As shown in Figure S7, substantial fairness improvements can be achieved by removing a small fraction of weights from each layer of the BERT-base model under certain configurations. This suggests that the optimal masking ratio may vary depending on the dataset and experimental setup.

D Analysis

D.1 Relationship of three types of masks in Dual Filter

The relationships between the ablation ratio of the three types of masks and the choice of k are shown

Methods	Hyparam	Values	Note
CONGATER	ω	$0.2, 0.4, 0.6, 0.8, 1.0$	*gate sensitivity*
MODDIFFY	`fixmask_pct`	$0.1, 0.3, 0.5, 0.7, 0.9$	*weight mask cutoffs*

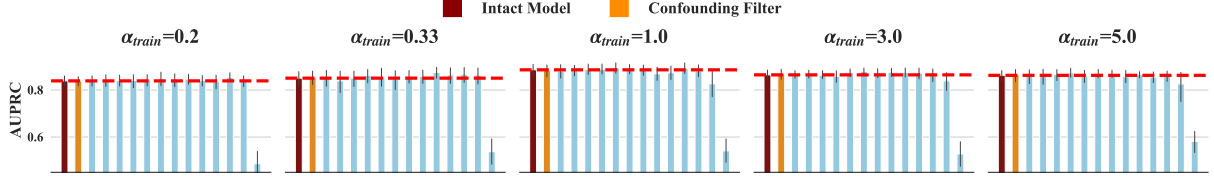Table S3: Hyperparameters used in baseline experiments

Figure S4: ECF filtering with 5% masking rate across different confounding shifts on DB dataset.
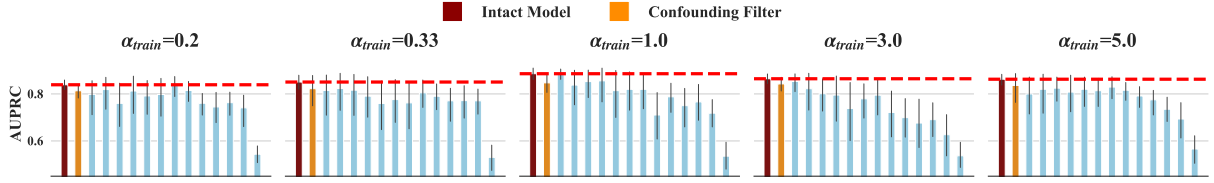

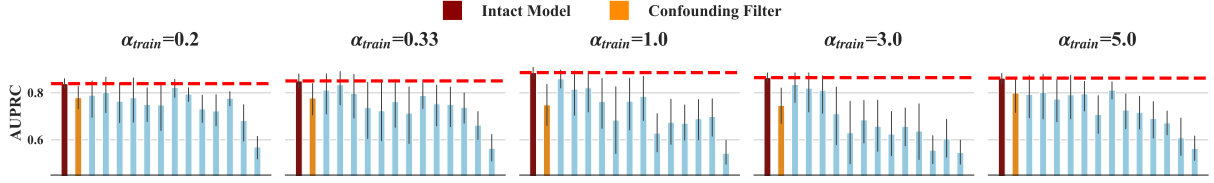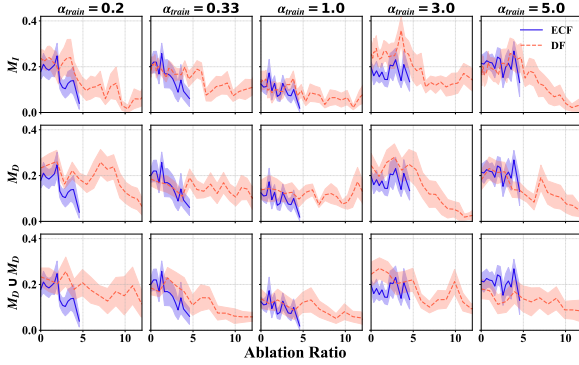Figure S5: ECF filtering with 25% masking rate across different confounding shifts on DB dataset.

Figure S6: ECF filtering with 35% masking rate across different confounding shifts on DB dataset.

Figure S7: Comparison of ECF with 5% masking ratio and DF on ΔFPR against ablation ratio on DB dataset.

in Figure S8. As we tune k to increase the coverage of active parameters in the model, the size of M_D first grows then reaches its peak at around $k = 40$ and then falls back to zero, while the size of M_I keeps increasing.

D.2 Entanglement Analysis

While the aim of this work is to eliminate gender confounding effects from the model's dementia detection capability, there is a possibility that the weights associated with dementia and gender become entangled during the learning process (i.e., same weights responsible for both - gender and dementia encoding). To investigate this, we record the
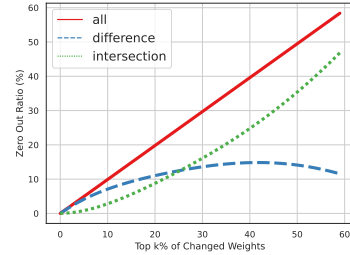

Figure S8: Ablation ratio by each mask against the total masking ratio

change matrices for all layers in the network during the Dual Filter training process. We then conduct an analysis of the similarity between the change matrices from the fine-tuned dementia model and those from the fine-tuned gender model. For similarity measurements, we utilize the Jaccard Index to quantify the similarity between the two input matrices, which is defined as:

$$J(U, V) = \frac{|U \cap V|}{|U \cup V|}$$

To prepare the input, 85% percentiles of the two change matrices are calculated and then the percentile values are used to binarize each of the matrices. Figure S9 to S13 demonstrates the barplot from six of the tracked weight matrices at each layer, with different α_{train} configurations. From

the plots we can observe that at lower encoder layers, the similarity between dementia model and gender model concentrates on the attention block, especially W_V and W_O. As we move up to the upper layer, the FFN block starts to display more similarity and jumps up at 12th layer. Similar patterns are also observed in other α_{train} configurations. This result indicates the fine-tuned model stores information dynamically through the whole network and shift the storage at different layers. This finding also aligns with other work (Wei et al., 2024) where weights entanglement are assessed with a larger model and different tasks.

Figure S9: Jaccard Index for each of the tracked matrix in Dual Filter ($\alpha_{train} = 0.20, \alpha_{test} = 5.0$)
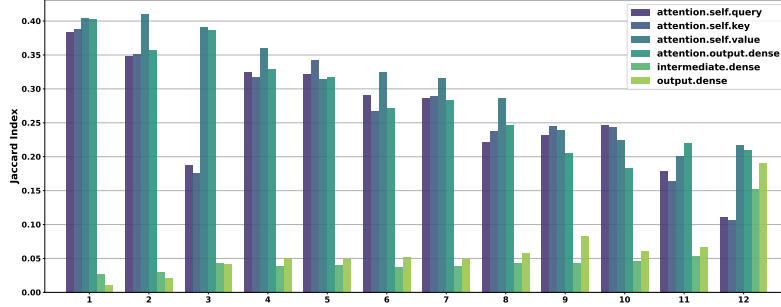

Figure S10: Jaccard Index for each of the tracked matrix in Dual Filter ($\alpha_{train} = 0.33, \alpha_{test} = 3.0$)
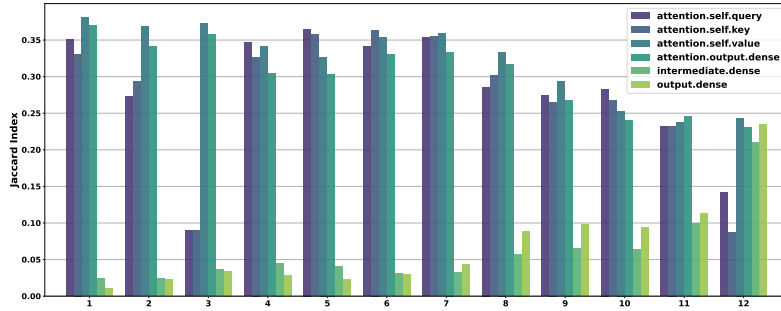

Figure S11: Jaccard Index for each of the tracked matrix in Dual Filter ($\alpha_{train} = 1.0, \alpha_{test} = 1.0$)
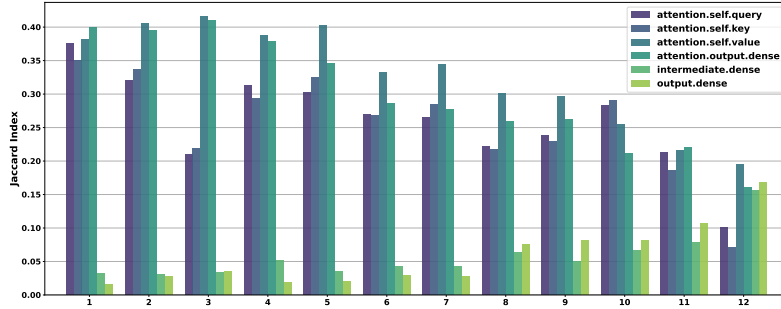

Figure S12: Jaccard Index for each of the tracked matrix in Dual Filter ($\alpha_{train} = 3.0, \alpha_{test} = 0.33$)
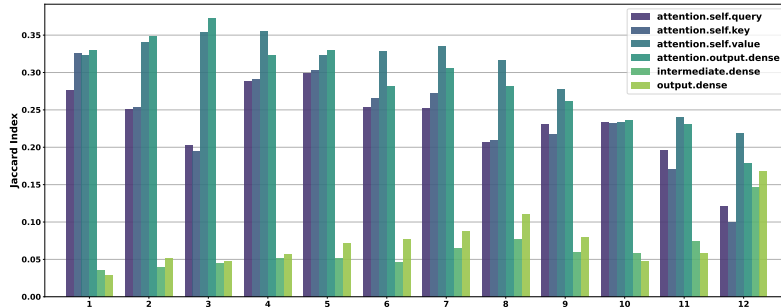

Figure S13: Jaccard Index for each of the tracked matrix in Dual Filter ($\alpha_{train} = 5.0, \alpha_{test} = 0.20$)