ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240769

Transfer Learning Can Introduce Bias

Parisa Salmani[a,*] **and Peter R. Lewis**[a]

[a]Ontario Tech University

Abstract. Transfer learning involves leveraging knowledge gained from solving one task and then using that knowledge to improve performance and reduce subsequent training time on a different but related task. Despite its advantages, recent attention has been directed towards a critical concern relating to the fairness of models trained with transfer learning. A previous study has demonstrated that transfer learning can preserve biases (that are intentionally planted) from the source task, transferring them to the target task. In this paper, we question a different but equally critical problem: whether transfer learning can introduce new biases or lead to greater biases in the target task compared to models trained from scratch. Our investigation reveals that transfer learning has the potential to introduce varying degrees of bias in the target task that were not originally present in the source task. Specifically, in an Alzheimer's Disease classification task, we show that the use of transfer learning introduces greater bias with respect to sex and age, compared to an equivalent non-transfer learning approach and a simpler model, both trained from scratch and almost as accurate. These findings underscore the need for a comprehensive understanding of the inherent limitations and risks associated with the application of transfer learning, particularly in high-risk applications, e.g. healthcare. This result also suggests the need for further research into how and when transfer learning introduces and amplifies bias.

1 Introduction

In recent years, machine learning has seen widespread adoption, yet some significant challenges persist such as the demand for an extensive volume of labeled data [56] or the subsequent time for training the model [40]. Overcoming these limitations is crucial for the advancement of machine learning applications. One of the strategies to address these challenges has been the employment of transfer learning [56]. This methodology seeks to mitigate the need for large labeled data by leveraging knowledge acquired from prior experiences [33]. Transfer learning represents a paradigm shift in machine learning, where models are pre-trained on large datasets for a specific task and then fine-tuned for a new task with a limited set of labeled data. This method is becoming more popular in today's scientific community since it allows developers to quickly build models for tasks that have less data.

The concept of transfer learning bears resemblance to a psychological phenomenon, whereby prior learning experiences influence the subsequent acquisition of related knowledge or skills [32, 3]. For instance, an individual proficient in playing the violin would likely find it easier to learn a new instrument like the piano [33]. Similarly, a pre-trained model tasked with classifying bicycles may exhibit higher performance in classifying motorcycles. This approach is most useful when data is scarce or expensive to obtain [23]. For example, in computer vision [9], transfer learning can be used to train models for specific tasks such as object detection or facial recognition by reusing the weights of a pre-trained model. Transfer learning has also been successfully used in large language models such as ChatGPT [48] and also in natural language processing [15] to improve text classification, sentiment analysis, and machine translation.

The fundamental idea of transfer learning is to train a model on a particular dataset (source task) and then either freeze the entire model or fine-tune only the last layers of the model on another task (target task) [33]. However, transfer learning models are pre-trained with specific datasets and retraining them only involves fine-tuning certain parameters on the last layers to adjust them to the new task. The question, therefore, is whether applying transfer learning to tasks with limited data might inadvertently introduce or amplify biases toward subgroups within the target task, stemming from the model's training on a different (source) task.

In the contemporary landscape of machine learning, while transfer learning has emerged as a widely employed technique, it is not without potential challenges, as highlighted by Salman et al. [39], who brought attention to the phenomenon of bias transfer. This phenomenon refers to the persistence of biases from the source task, even after adaptation, to a target task.

A critical aspect raised by Salman et al. prompts us to examine the scenario where the source task exhibits no discernible bias and appears to be well-performing. Yet, when the pre-trained model is employed for a different task, it manifests biases towards specific subgroups within the new task's data. It should be noted that while the related scope of work like [39] is acknowledged, the emphasis in this paper is the introduction of bias through transfer learning in specific scenarios, not bias transfer when the initial task is intentionally biased. This is a question that has not been previously addressed in the literature. This hypothesis can hold substantial significance, as there is a tendency to assume that a model performing optimally on the first task will seamlessly generalize when fine-tuned for and applied to a second task. Consequently, the requirement to scrutinize for bias in the latter context is often overlooked. This concern is magnified when the application of transfer learning is in critical domains such as healthcare. Given the potential consequences of biased predictions in such high-stakes settings, where bias could have a direct impact on people, a comprehensive understanding of bias in transfer learning becomes imperative. This paper explores the question: Can the utilization of transfer learning introduce biases for the target task?

* Corresponding Author. Email: parisa.salmani@ontariotechu.net

2 Related Work

2.1 Transfer Learning

Producing accurate predictions based on available information and data is crucial in many industries. To effectively handle data, employing machine learning and data analysis techniques is essential [10]. Deep learning, a widely employed machine learning approach, is capable of obtaining complex information out of huge amounts of data. It has achieved considerable success across various applications, notably in computer vision and natural language processing [11]. Despite the remarkable performance exhibited by deep learning models, they still encounter inherent limitations and drawbacks which need to be solved. Two prominent challenges are 1) The substantial demand for extensive training data, which may be unavailable, expensive, or difficult to collect [52], and 2) The considerable time required for training models, particularly when dealing with extensive datasets [22]. In addressing these challenges, various techniques have been developed. Among these methods, data augmentation is employed to alleviate data scarcity problems [5], while batch optimization strategies are applied to enhance training time [53]. Along with these approaches, transfer learning emerges as a prominent solution capable of simultaneously addressing both issues. Transfer learning utilizes knowledge gained from solving one task to enhance performance and accelerate training on a related task [33].

Transfer learning can be categorized into inductive [19] and transductive [4]. Another categorization is heterogeneous and homogenous transfer learning [51]. Additionally, there are various methodologies for implementing transfer learning, including instance transfer [13] and parameter transfer [25]. Several areas of study are closely related to transfer learning, such as semi-supervised learning [43]. Semi-supervised learning, which lessens the reliance on labelled data, still necessitates a substantial amount of data but may not require all labels. Moreover, concepts such as domain adaptation [34], multi-task learning [54], life-long learning [2], inductive transfer [6], knowledge transfer [44], and incremental learning [35] share some similarities with transfer learning. However, while there may be implications for those areas from our results, they are not the same and, as such, fall outside the scope of the current research.

Another aspect of transfer learning is the concept of negative learning [49]. This occurs when the model's performance on the target task can be adversely affected by weak connections between the source and the target task. It is important to note that negative learning, while can be perceived to be similar to bias, is a distinct concept that we do not explore in this research.

The applications of transfer learning are vast and varied, encompassing fields such as computer vision (image classification) [9, 28], natural language processing (text and sentiment classification) [36, 29, 47, 37], and recommender systems (movies, books, etc.) [8, 55]. The versatility of transfer learning makes it a valuable tool in numerous domains.

Transfer learning has found widespread application in healthcare, with numerous studies conducted in recent years [12, 14, 17, 50, 21]. A common area of focus is Alzheimer's disease (AD) classification, which is the focus of this paper. Using deep learning for healthcare encounters several challenges. The need for a large amount of labeled training data is a major obstacle, especially in fields like medical imaging, where obtaining annotated data is costly and restricted across institutions due to ethical concerns. Additionally, training deep networks with a large number of images demands significant computational resources, raising feasibility and efficiency concerns. A popular solution involves fine-tuning pre-trained deep networks, especially Convolutional Neural Networks (CNNs), through transfer learning. This method tackles challenges related to limited data and computational resources, providing a more efficient means of applying deep learning to healthcare tasks.

In a paper [21] focusing on Alzheimer's detection, Hon and Khan proposed a transfer learning-based method to detect AD from MRI images. They tested two popular architectures, namely VGG16 [41] and Inception V4 [42], using transfer learning with pre-trained weights from the model trained on the ImageNet dataset and fine-tuning them on the MRI images. They achieved comparable results with a small number of training images compared to the other approaches. Additionally, they employed an entropy-based technique to select the training dataset, ensuring it represented the most information within a small set. They claimed that their method provided performance comparable to other methods, despite having a training set many times smaller.

Leveraging their entropy-based technique and using their study as a starting point, we are employing the same dataset, architecture of the model they used (VGG16), and same application to examine the bias introduced by transfer learning. While many researchers and studies are employing transfer learning using different foundation models and datasets, we intentionally focused on this specific, well-cited study [21] as a case study to demonstrate that this highly popular approach is not always optimal or neutral.

2.2 Bias in Machine Learning

As artificial intelligence (AI) systems and their applications become increasingly prevalent in our daily lives, the consideration of fairness in AI has become a significant aspect in the design and engineering of such systems. AI systems have also been utilized in sensitive environments such as medicine [46] to make critical and life-altering decisions, underscoring the importance of ensuring that these decisions do not include bias towards specific groups or populations. Recently, there has been growing research in traditional machine learning and deep learning, addressing such challenges in various subdomains [1, 31, 26].

To ensure that the algorithms are fair in machine learning, it is vital to understand the concepts of bias and fairness within their application. Verma and Rubin's [45] research synthesizes the most significant definitions of algorithmic fairness for classification problems, illustrating their rationales and applications through a single comprehensive case study. Among the 20 definitions of fairness that they explained, the commonly recognized ones include group fairness or statistical parity [16], equalized odds [20], and fairness through awareness [16]. While these definitions are relevant in certain situations, another important definition to consider – particularly in medical diagnosis – is accuracy parity [7]. This definition becomes significant when the probability of a subject being predicted as either positive or negative is equally important. Essentially, a classifier is fair within this definition if it ensures equal prediction accuracy for both positive and negative groups.

Accurate diagnosis is crucial in healthcare, and errors like false negatives and false positives can lead to significant consequences. A false negative may result in delayed treatment and interventions for the patient, while a false positive can cause unnecessary stress, anxiety, and undesired financial and legal implications both for the individual and their family. Moreover, incorrect medication may be prescribed, affecting the patient's health. Therefore, in healthcare, it is essential to strive for accurate predictions in both positive and negative cases to ensure the best possible outcomes for patients.

In this particular application, as in healthcare, accuracy and the ability to correctly assign patients to their respective groups are crucial. It is equally desirable to accurately classify both patients with Alzheimer's disease and patients without the disease across all sex and age groups. Therefore, for this study, we adopt the definition of accuracy parity and use it as our bias metric.

2.3 Bias in Transfer Learning

Existing literature primarily focuses on enhancing the performance of transfer learning models when applied across different domains. However, real-world scenarios often exhibit differences between the source and target domains. In a study [27], Li highlighted the Domain Class Imbalance (DCI) issue, where class samples in two domains have different ratios. For instance, in a binary classification problem, a 50/50 (Pos/Neg) balance might exist in the source task, but the target task might have a 30/70 ratio. In such cases, if a classifier predicts all samples as negative cases, it can still achieve an accuracy of 70%. However, in critical fields like medicine, accurate detection of positive and negative cases is crucial. While many studies have reported an average improvement in accuracy and F1 score across all classes, few have examined the improvements for each class, particularly for rare ones. This work aims to provide an analysis of the robustness of deep transfer learning models in text classification tasks under a domain class imbalanced setting. Similar to the idea of the above-mentioned paper, our objective is to investigate whether the overall accuracy of a pre-trained model can potentially mask discrimination towards a specific subgroup, leading to lower accuracy for that subgroup compared to others.

As introduced in Section 1, in another study [39], Salman et al. showed that biases that exist (and intentionally planted) in pre-trained models trained on a source task often persist even after fine-tuning on the target tasks. Importantly, these biases can persist even when the target dataset used for fine-tuning does not exhibit such biases. To substantiate this, they deliberately planted bias into the source task and observed that the bias would remain in the target task even after fine-tuning. For example, by creating datasets that amplify certain spurious correlations, the authors demonstrated that models trained on these biased datasets continue to exhibit sensitivity to these correlations in the target dataset. Through a blend of synthetic and natural experiments, they have shown that bias transfer not only occurs in realistic scenarios but also can happen even when the target dataset is explicitly de-biased.

In this paper, inspired by their results, we question a different, yet important problem: whether the transfer learning approach can inadvertently introduce bias towards the second task, even if the source task does not contain bias.

3 Problem Statement

Transfer learning, a prevalent approach in machine learning, involves utilizing pre-trained models that have been initially trained on huge datasets for a specific task. The primary aim is to transfer the acquired knowledge from the pre-trained model to new tasks, thereby speeding up the training process and enhancing performance with the scarce dataset. However, an emerging concern is the potential introduction of bias when employing pre-trained models on tasks for which they were not originally designed. Transferred models can show bias towards certain subgroups or amplify existing bias existed in the second task compared to non-transfer learning models trained from scratch. For instance, in the context of healthcare, the use of transfer learning models could lead to displaying bias towards demographic subgroups, as a consequence of the previous learning task upon which the model was initially trained. However, using the exact same model and architecture which was trained from scratch (only on the second dataset) would not demonstrate such bias.

The machine learning models are trained to excel in the prediction of their primary task by capturing the patterns and features relevant to that specific task using the dataset. This specificity, while advantageous for the primary task, may inadvertently lead to complications when these models are transferred to secondary tasks. The knowledge encoded within pre-trained models might not align with the requirements of the new task, potentially introducing bias due to mismatches between the primary and secondary datasets.

Retraining the pre-trained model on the secondary task aims to address the mismatch between the datasets and adjust the model to the secondary task. However, retraining predominantly involves adjusting the model's weights (retraining the model with the new dataset to update the parameters), which might not fully rectify the misalignment between the primary and secondary tasks. Consequently, while retraining aligns the model with the second task to some extent, biases toward the second task's subgroups may persist and remain hidden within the acceptable accuracy of the transferred model.

To demonstrate the evidence of bias introduced by transfer learning, we analyze a previous study [21] using their pre-trained transfer learning model, fine-tuned on the target task, and compare its performance with models with the same architecture trained from scratch. To ascertain that this bias is towards specific subgroups, we can observe the differences in accuracy among these subgroups categorized based on sensitive attributes in the dataset. The following sections look into the bias definition and the experiments.

4 Bias Metric

To explore the notion of bias in transfer learning, we focus on biases that may emerge towards different subgroups. The primary objective is to utilize a robust and quantifiable metric to assess the presence of bias in transfer learning models.

It is worth mentioning that there is no universal bias metric and the most appropriate one (or several) depends on the application. In light of this, the role of scientists is to provide accurate information where there is consistency of biased behaviour in order that decision-makers, together with stakeholders and those affected, can exercise informed judgment concerning potential harm associated with such biases.

Among the various bias metrics used to evaluate model fairness with respect to sensitive attributes, two prominent ones are Demographic Parity [18] and Equalized Odds [20]. Demographic Parity seeks to ensure that all subgroups receive a proportionate share of positive outcomes, while Equalized Odds aim to provide each subgroup with positive outcomes at equal rates.

In the context of our experiments, we have chosen to focus on the metric known as Accuracy Parity [45, 18]. This choice aligns with our specific research objectives. While metrics like true positives, true negatives, false positives, and false negatives are undeniably important, our primary concern lies in preventing the model from becoming excessively biased towards one subgroup at the expense of another, thus leading to a situation where one group experiences a disproportionately low accuracy rate, even if the overall accuracy appears satisfactory. In this scenario, while the overall accuracy may seem acceptable, it masks the accuracies of individual subgroups, causing hidden bias in the results.

In the context of transfer learning in our healthcare application, bias emerges as the model's differences in accuracy among the subgroups, result in disparities in performance. To quantify this bias in our investigation, we employ a straightforward definition based on the absolute difference between the highest and lowest accuracy levels observed among subgroups. This approach allows us to highlight meaningful differences while providing an interpretable metric.

Our definition of bias encapsulates the disparity between the highest accuracy achieved among subgroups and the lowest accuracy within those subgroups. While a simplistic assumption suggests that the maximum bias would be 100%, occurring when one subgroup achieves 100% accuracy and the other 0%, this scenario is improbable unless the model intentionally contradicts its learned patterns. A more reasonable minimum accuracy arises when the model fails to learn for one subgroup, resulting in a 50/50 prediction guess and 50% accuracy in a binary situation. This quantification of bias spans from 0, denoting the absence of bias aligning with our central objective, to 50 as the upper limit. However, this is also implausible as achieving 100% accuracy for one subgroup means learning everything and achieving 50% for another implies that the model may not be learning anything useful for the other subgroup within the same dataset. If the model fully masters one subgroup and simultaneously learns nothing for another, it suggests a lack of any common knowledge between the concepts relevant to these subgroups. In practical terms, this means the model is treating these subgroups as disjoint tasks, which is typically not the case in a real-world dataset. Further exploration of this topic will be deferred for future work due to its complexity, necessitating additional time for in-depth study.

In practical terms, the assessment of statistical fairness metrics, as delineated by Verma and Rubin [45], establishes a threshold for bias evaluation. A bias magnitude below 3% is categorized as minor, while a deviation exceeding 6% is deemed a major difference. Consequently, any bias surpassing the 3% threshold is considered noteworthy. A comprehensive examination of the bias threshold is outside the scope of this paper and will be deferred to future work.

Accuracy Parity is defined as follows [18]:

$$P(C = Y|A = a) = P(C = Y|A = b) \tag{1}$$

Therefore, the bias metric can be defined as:

$$Bias = |P(C = Y|A = a) - P(C = Y|A = b)| \tag{2}$$

In the context of bias definition, a and b represent distinct subgroups, and the outcome variable is denoted by Y.

This measure is a defined bias metric aimed at finding the difference in accuracy across subgroups to avoid fluctuating accuracy and uncover hidden disparities within a seemingly high overall accuracy score. This metric is commonly used in many studies, such as a study by Kärkkäinen and Joo [24] to discover biases within face datasets. The definition provided above typically applies to situations involving binary distinctions, such as sex subgroups. However, when dealing with multiple subgroups, we can describe it as the difference in accuracy between the subgroup with the highest accuracy and the subgroup with the lowest accuracy:

$$Bias = max_{sa} - min_{sa} \tag{3}$$

Where max_{sa} and min_{sa} are the maximum and minimum subgroup accuracies, respectively.

Throughout our experiments, we applied the proposed bias calculation method to measure the bias across our dataset. The results showcase the effectiveness of this methodology in detecting and quantifying bias, providing insights into the model's subgroup-specific performance.

5 Methodology

Transfer learning is widely applied across various domains, with healthcare being no exception. Specifically, it has earned significant attention in the realm of Alzheimer's disease classification, an important and life-impacting area of healthcare. As an example, Hon and Khan [21], in their study, demonstrated that transfer learning can enhance accuracy and efficiency in Alzheimer's disease classification while requiring less time training. Our experiments are built directly upon their experiments, with the same transfer learning approach, model architecture and dataset. This is to ensure comparability and validity compared to an existing approach.

Our study focuses on assessing the presence of bias in this high-stakes application. To conduct our investigation, we utilize the preprocessed OASIS dataset, which was originally introduced in the Journal of Cognitive Neuroscience [30].

This dataset comprises 3D scan images of patients' brains. Hon and Khan [21] employed an entropy-based sorting mechanism to select the most informative images from the axial plane of each 3D scan across 200 patients and introduced the preprocessed dataset in their paper. This selection process resulted in a total of 6,400 training images. Among these, 3,200 were associated with Alzheimer's disease (AD), while the remaining 3,200 represented non-Alzheimer's cases. This dataset forms the basis of our investigation into the potential introduction of bias in transfer learning for Alzheimer's disease classification.

In this study, our primary objective is to assess the potential introduction of bias through the utilization of transfer learning. To achieve this, we employed three distinct models and conducted a comparative analysis of the degree of bias present in each. The first model denoted as VGG16, underwent initial training on the ImageNet dataset [38], followed by fine-tuning on the target dataset. The second model, also VGG16, possessing an identical architecture to the previous model, was trained entirely from scratch on the brain images dataset, thereby allowing for an authentic examination of bias in both transferred and non-transferred models. The third model, designed with a simpler architecture akin to VGG16 but featuring fewer layers and parameters, was considered to address the potential challenge of VGG16's size being impractical for a relatively small dataset like OASIS when trained from scratch. Consequently, a logical approach involved deploying a more straightforward model and subsequently comparing its results with the two aforementioned models.

Predictions were obtained from all three models, and bias was quantitatively assessed using the defined metric, as outlined in equation 3. This comprehensive approach enables a thorough examination of bias across different model architectures, revealing the impact of transfer learning on bias introduction.

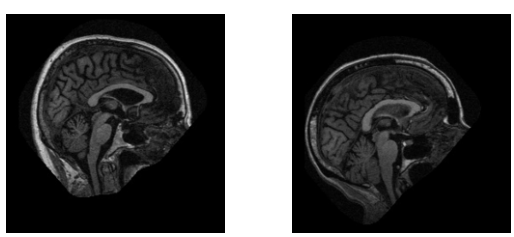

Figure 1: Sample Brain Images from the Target Task Dataset [30, 21]

In light of the dataset characteristics, we identified two pertinent sensitive attributes—age and sex—to systematically investigate the fairness of our model with respect to these demographic factors. The dataset was split twice to facilitate a nuanced exploration of potential biases. Firstly, the sex attribute prompted the creation of two distinct subgroups: one comprising female individuals and the other, male individuals. Secondly, to capture the age-related biases, the dataset was divided into 13 subgroups based on age, each encompassing a 5-year range. The rationale behind choosing a 5-year range was a pragmatic balance; a single year per subgroup would yield an unwieldy number of subgroups, while a 10-year range would result in overly broad groupings.

These considerations bear particular significance, particularly in the context of healthcare applications, where ensuring model fairness is paramount. It is imperative to ascertain that the model exhibits unbiased behavior across diverse demographic categories, as discrepancies in performance may manifest even when the overall accuracy appears high. This approach allows for the identification of potential biases that might be obscured by an aggregated assessment, resulted in any disparities linked to sex or age, critical considerations in healthcare equity.

6 Experiments

In this study, we aim to assess the extent of bias introduced by a transferred model across two key demographic attributes: age and sex. The investigation focuses on determining whether the transferred model imparts new biases towards specific groups within these demographic categories. To facilitate a comprehensive analysis, three distinct models are employed for each sensitive attribute, each sharing architectural similarities.

6.1 Attributes

6.1.1 Sex

The dataset under consideration comprises 200 subjects (individuals), evenly divided into two cohorts: 100 with Alzheimer's disease and 100 without. Each subject has 32 images of their brain resulting in 6400 images in total. In this section, we split the entire dataset based on sex, aiming to discern differences in model accuracy between male and female subgroups. Despite potential imbalances in the distribution of these two subgroups, we maintain consistency in the evaluation process across all three models, thereby comparing their performance under the same conditions.

By adopting this approach, we deliberately overlook the impact of data imbalances on bias, allowing for a focused analysis of the models' impact on accuracies in each sex category. This decision ensures a rigorous examination of the models' performance, enabling us to evaluate biases introduced by the models only, rather than confounding effects arising from subgroup imbalances. In this study, our focus is yet not on mitigating bias or providing strategies for bias reduction in transfer learning. Instead, we aim to compare the bias levels in two distinct approaches (transfer learning vs. non-transfer learning) solely based on the chosen model for experimentation. The goal is to understand whether opting for the transfer learning approach can influence the bias observed in the predicted results.

To obtain the sex subgroups, we read individuals' data from a CSV file that came with the brain images dataset, and which categorized them as either male or female based on their sex. Subsequently, we assess the average accuracy of each model within each sex subgroup, allowing us to measure bias by comparing the accuracy discrepancies between the male and female groups.

6.1.2 Age

In parallel with the sex subgroup analysis, the dataset, including individuals with the age range of 18 to 96, has been further split based on age, dividing it into 13 subgroups with a range of 5 years for each subject. It is worth mentioning that the initial seven subgroups were excluded from the experiments due to either the absence of data or the presence of only one subject, categorizing them as outliers. This exclusion was deemed necessary to maintain the robustness and reliability of the subsequent analyses.

Similar to the approach taken with sex subgroups, any distributional differences within the age subgroups are disregarded, as these variations are consistent across the evaluations of all three models. By adopting a uniform methodology, the study aims to ensure that comparisons between models remain unaffected by potential imbalances in subgroup distributions. This approach allows for a focused examination of the models' performances within distinct age categories, free from the confounding effects of outlier subgroups or distributional disparities.

Similarly, for the age attribute, we sort subjects' brain images into relevant subgroups after reading their data. We then calculate the average accuracy within each age group and find the difference between the minimum and maximum accuracy across all subgroups as the bias. This process provides a comprehensive understanding of the model's performance across different age brackets, enabling a thorough assessment of potential biases.

6.2 Models

6.2.1 Transfer Learning Approach: Pre-trained VGG16

In our experiment, following Hon and Khan's [21] work, we employ the VGG16 model, a widely used choice for transfer learning. Developed by Oxford's Visual Geometry Group, this 16-layer neural network serves as our starting point. Following their methodology in [21], we initialize the model with weights from the Keras library which uploads the weights of the VGG16 model pre-trained on the ImageNet dataset. Re-implementing the referenced work, we fine-tune the last three layers of VGG16 using our brain dataset, constituting the second task in our study.

The final three layers are configured as follows. Firstly, a sequential model is created, featuring a flattened layer designed for one-dimensional input. Subsequently, a dense layer with 256 units and ReLU activation is added, followed by a dropout layer with a 0.5 dropout rate. Another dense layer, equipped with a single unit and sigmoid activation for binary classification, is then incorporated. The model is compiled utilizing the RMSprop optimizer, and the binary cross-entropy loss. The experiment is conducted for 150 epochs, employing a batch size of 40 and utilizing validation data for evaluation. Results were obtained using 5-fold cross-validation, with an 80-20 split between training and testing. To assess performance comprehensively, the model undergoes 10 runs, with the final accuracy and validation accuracy recorded and stored for further analysis.

Throughout these runs, we record subgroup accuracy based on sensitive attributes. We then calculate accuracy differences between these subgroups, and the mean of these differences among the 10 runs serves as the bias metric. This straightforward approach ensures a thorough evaluation of the model's performance, highlighting potential biases within our brain dataset.

6.2.2 Training From Scratch: VGG16

To ensure a fair comparison between the results concerning bias introduced in transfer learning and those from the non-transfer learning approaches, as is the purpose of our study, we employ the VGG16 architecture as an identical architecture to the previous model. However, unlike the transferred model that utilized pre-trained weights from ImageNet, the VGG16 model in this experiment is trained from scratch while keeping other parameters the same as the previous model. The model weights are randomly initialized from a normal distribution, maintaining other parameters consistent with the pre-trained VGG16. The experimental setup includes 10 runs, each has 150 epochs with a batch size of 40, ensuring a direct and fair comparison between the two models.

6.2.3 Training From Scratch: Baseline

Our baseline model, inspired by the Keras library tutorial, is a simplified Deep Convolutional Neural Network (CNN) with 8 layers, trained from scratch on the brain image dataset. Due to its simplicity and fewer parameters, this model requires additional epochs to effectively learn features. By allocating just twice the training time, it can attain a reasonable level of accuracy. Therefore, in 10 training runs, each 300 epochs with a batch size of 40, we compare the results of this simpler model with the other two models.

The model is constructed sequentially, featuring three convolutional layers (32, 32, and 64 filters) with ReLU activation and (2, 2) max pooling. Post-convolution, a flattening layer transforms the 3D data into a 1D vector. A dense layer with 64 units and ReLU activation are added, accompanied by a dropout layer (rate: 0.5) to prevent overfitting. The output layer, designed for binary classification, comprises a single unit with sigmoid activation. The model is compiled with SGD optimizer (learning rate: 0.005) and binary cross-entropy loss, standard for binary classification.

Despite setting a random seed for each model to ensure reproducibility, utilizing CUDA introduces a challenge. When operations are executed on a GPU, some outputs may become non-deterministic. This unpredictability arises from the parallel nature of GPU operations, where the order of execution is not always guaranteed.

7 Results

Table 1 highlights a substantial degree of bias in the transfer learning model compared to the models trained from scratch. Despite the bias for sex attributes being relatively small, it remains about 3 times higher compared to the other two models. Notably, the bias associated with the age attribute is 4-5 times more pronounced than observed in the other models, underscoring a discernible bias toward this attribute—a level of bias which is deemed noteworthy in [45].

Bias and overall accuracy are correlated; higher accuracy inherently tightens the lower bound of bias, whereas lower accuracy expands the upper bound of bias, as per mathematical principles. Notably, the non-transfer learning VGG16 model that was trained from scratch and the baseline model achieve nearly identical accuracy to the pre-trained model but exhibit significantly less bias, showing the fact that the transfer learning approach is introducing some new level of bias in this application.

In this specific application, we find that using transfer learning can achieve satisfactory results in 100 epochs. Alternatively, training the same model from scratch also yields good results, taking only an additional 50 epochs, making it 150 epochs. The choice between the two approaches lies between time and bias trade-off. Training the last three layers of the VGG16 model using the transfer learning approach is approximately 7-8 times faster than training all 16 or even 8 layers from scratch. However, the VGG16 model trained entirely from scratch and the baseline model, our non-transfer learning models, exhibit 3-4 times less bias compared to its transfer learning counterpart. Therefore, the trade-off in this scenario lies between the time invested and the level of bias in the model.

Figure 2 shows the accuracy scores across the age-based subgroups, with an obvious bias between the first two subgroups in the pre-trained model, as indicated by both the mean and standard deviation values. Notably, this bias is negligible in the case of the two other models. To validate the significance of these observations, a thorough analysis of the results is conducted by examining the corresponding P-values using the T-test approach. It is worth mentioning that all the assumptions of the T-test are satisfied. These assumptions including the independence of results (satisfied by the nature of experiment as there are different training runs) and the normal distribution of the results (verified using the Shapiro-Wilk test) are all satisfied, and therefore, we are able to use this test to measure the significance of our results.

For this analysis, the null hypothesis posits that the observed variations in minimum and maximum accuracy, along with the bias across, are merely products of the randomness of the model training. Employing a T-test on the obtained results yields a P-value of $0.00090 < 0.05$. The rejection of the null hypothesis implies that the differences in accuracy and the observed bias are not attributable to randomness, confirming the significance of our findings.

8 Discussion

This paper examines bias levels in a transfer learning approach, emphasizing the potential for the introduction of bias to a target task, even when the source task and its dataset lack such bias. Motivated by the prevailing assumption that transfer learning has minimal drawbacks, we highlight the associated high risks. It is important to highlight that this research indicates the potential risks that implementing transfer learning can have.

This contribution of this study is therefore to show that the introduction of bias can occur simply by using transfer learning. However, the full scope of this phenomenon remains uncertain. We yet need to determine the extent of its impact on other applications and the broader ubiquity of this phenomenon. We propose practitioners assess the bias introduced by transfer learning using an application-specific bias definition and determine its significance in consultation with those affected, based on the potential form of the harm. This assessment empowers practitioners to decide whether to opt for transfer learning or pursue model training from scratch. The recommendation gains particular significance in critical domains like healthcare and justice decision-making, where biases can yield profound consequences. Therefore, practitioners should conscientiously weigh biases aligned with the application's requirements before adopting transfer learning.

9 Conclusion

In conclusion, our study challenges the common belief that transfer learning is uniformly advantageous in data-limited scenarios. Instead, we reveal a notable bias through our experiments toward

Bias Analysis			
Model	Validation Accuracy (%)	Sex Bias (%)	Age Bias (%)
Transfer Learning VGG16	92.30±0.17	**0.9517 ± 0.277**	**5.2524 ± 1.981**
VGG16 Trained From Scratch	91.73±0.09	0.3409±0.102	0.990±0.167
Baseline	91.81±1.7	0.3995±0.346	1.354±0.837

Table 1: The table indicates that the model exhibiting the highest degree of bias among the three is the one employing transfer learning. Despite non-transfer learning VGG16 having the same architecture as the transfer learning model, the bias for the model trained from scratch is three to five times lower. Similarly, the baseline model shows approximately three times less bias in both sex and age compared to the pre-trained transfer learning model.

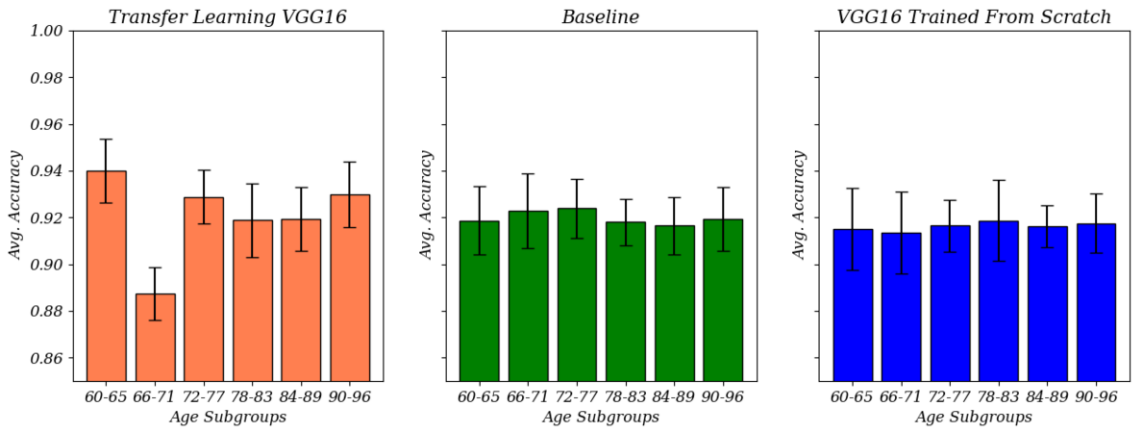

Figure 2: As demonstrated in the figure, a substantial accuracy gap exists between the first two subgroups, categorized by age, in the transfer learning approach. On the contrary, the accuracies of the subgroups for the other two non-transfer learning models remain consistent and nearly identical. This leads us to the conclusion that the subgroup with lower accuracy is masked within the transfer learning approach's overall high accuracy.

the second task when using transfer learning compared to non-transfer learning approaches. Our experiments show that a pre-trained VGG16 model transferred from a different task displays accuracy disparity across subgroups of data in the second task, with this bias being at least three times greater than models entirely trained from scratch as the non-transfer learning approach. Ultimately, the impact of this bias depends on the specific application of transfer learning. Practitioners should recognize the potential harm and inform end-users of existing biases, underscoring the need for careful consideration in the deployment of transfer learning models.

10 Future Work

In future research, it would be beneficial to explore other applications of transfer learning, to ascertain if the biases it can generate occur elsewhere. Furthermore, evaluating alternative notions of bias may provide insight into the performance of different methodologies in light of these varying bias definitions. Exploring other base models that serve as foundations for transfer learning would also be valuable. Such an exploration could help us understand whether employing different model types with transfer learning can also lead to bias towards the second task. Additionally, a thorough examination of the correlation between overall accuracy, subgroup distributions, and the accuracy parity metric observed in each model could lead to the establishment of a concrete correlation formula.

References

[1] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, R. J. Bandara, and K. Michael. Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, 144:201–216, 2022.

[2] H. Ammar, E. Eaton, J. Luna, and P. Ruvolo. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. IJCAI International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, 2015.

[3] L. Argote, P. Ingram, J. M. Levine, and R. L. Moreland. Knowledge transfer in organizations: Learning from the experience of others. *Organizational Behavior and Human Decision Processes*, 82(1):1–8, 2000.

[4] A. Arnold, R. Nallapati, and W. W. Cohen. A comparative study of methods for transductive transfer learning. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 77–82, 2007.

[5] M. A. Bansal, D. R. Sharma, and D. M. Kathuria. A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Comput. Surv.*, 54(10s), sep 2022.

[6] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, Mar. 2000.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.

[8] N. Biadsy, L. Rokach, and A. Shmilovici. Transfer learning for content-based recommender systems using tree matching. In *Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013. Proceedings 8*, pages 387–399. Springer, 2013.

[9] A. Brodzicki, M. Piekarski, D. Kucharski, J. Jaworek-Korjakowska, and M. Gorgon. Transfer learning methods as a new approach in computer

vision tasks with small datasets. *Foundations of Computing and Decision Sciences*, 45:179–193, 10 2020.

[10] I. S. Candanedo, E. H. Nieves, S. R. González, M. T. S. Martín, and A. G. Briones. Machine learning predictive model for industry 4.0. In L. Uden, B. Hadzima, and I.-H. Ting, editors, *Knowledge Management in Organizations*, pages 501–510, Cham, 2018. Springer International Publishing.

[11] X.-W. Chen and X. Lin. Big data deep learning: Challenges and perspectives. *IEEE Access*, 2:514–525, 2014.

[12] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

[13] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 193–200, New York, NY, USA, 2007. Association for Computing Machinery.

[14] M. De Bois, M. A. El Yacoubi, and M. Ammi. Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine*, 199: 105874, 2021.

[15] C. B. Do and A. Y. Ng. Transfer learning for text classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness, 2011.

[17] A. Farhadi, D. Chen, R. McCoy, C. Scott, J. A. Miller, C. M. Vachon, and C. Ngufor. Breast cancer classification using deep transfer learning on structured healthcare data. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 277–286, 2019.

[18] A. Fraenkel. Fairness & algorithmic decision making: Lecture notes for ucsd course dsc 167, 2020. URL https://afraenkel.github.io/fairness-book/intro.html.

[19] J. Garcke and T. Vanck. Importance weighted inductive transfer learning for regression. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 466–481, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[20] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[21] M. Hon and N. Khan. Towards alzheimer's disease classification through transfer learning, 2017.

[22] M. L. Hutchinson, E. Antono, B. M. Gibbons, S. Paradiso, J. Ling, and B. Meredig. Overcoming data scarcity with transfer learning, 2017.

[23] M. L. Hutchinson, E. Antono, B. M. Gibbons, S. Paradiso, J. Ling, and B. Meredig. Overcoming data scarcity with transfer learning, 2017.

[24] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[25] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, New York, NY, USA, 2004. Association for Computing Machinery.

[26] S. Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18, page 14–16, New York, NY, USA, 2018. Association for Computing Machinery.

[27] I. Li. Detecting bias in transfer learning approaches for text classification, 2021.

[28] X. Li, Y. Grandvalet, F. Davoine, J. Cheng, Y. Cui, H. Zhang, S. Belongie, Y.-H. Tsai, and M.-H. Yang. Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing*, 93:103853, 2020.

[29] A. Malte and P. Ratadiya. Evolution of transfer learning in natural language processing. *arXiv preprint arXiv:1910.07370*, 2019.

[30] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507, 09 2007.

[31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[32] T. J. Nokes. Mechanisms of knowledge transfer. *Thinking & Reasoning*, 15(1):1–36, 2009.

[33] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[34] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[35] R. Polikar, L. Upda, S. Upda, and V. Honavar. Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(4):497–508, 2001.

[36] S. Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[37] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18, 2019.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[39] H. Salman, S. Jain, A. Ilyas, L. Engstrom, E. Wong, and A. Madry. When does bias transfer in transfer learning?, 2022.

[40] A. Shrestha and A. Mahmood. Review of deep learning algorithms and architectures. *IEEE Access*, 7:53040–53065, 2019.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[43] P. Thomas. Semi-supervised learning by olivier chapelle, bernhard schölkopf, and alexander zien (review). *IEEE Transactions on Neural Networks*, 20:542, 01 2009.

[44] G. N. Thompson, C. A. Estabrooks, and L. F. Degner. Clarifying the concepts in knowledge transfer: a literature review. *Journal of Advanced Nursing*, 53(6):691–701.

[45] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[46] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. Mitigating bias in machine learning for medicine. *Commun. Med. (Lond.)*, 1(1):25, Aug. 2021.

[47] D. Wang and T. F. Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE, 2015.

[48] D.-Q. Wang, L.-Y. Feng, J.-G. Ye, J.-G. Zou, and Y.-F. Zheng. Accelerating the integration of chatgpt and other large-scale ai models into biomedical research and healthcare. *MedComm – Future Medicine*, 2 (2):e43, 2023.

[49] Z. Wang, Z. Dai, B. Poczos, and J. Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[50] K. Weimann and T. O. Conrad. Transfer learning for ecg classification. *Scientific Reports*, 11(1), 2021.

[51] K. Weiss, T. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3, 05 2016.

[52] K. Weiss, T. M. Khoshgoftaar, and D. Wang. *Transfer Learning Techniques*, pages 53–99. Springer International Publishing, Cham, 2016.

[53] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2020.

[54] Y. Zhang and Q. Yang. An overview of multi-task learning. *National Science Review*, 5:30–43, 01 2018.

[55] L. Zhao, S. Pan, E. Xiang, E. Zhong, Z. Lu, and Q. Yang. Active transfer learning for cross-system recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1205–1211, 2013.

[56] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.