Exploring Bias in Pre-Trained CNNs: Fairness Assessment in Sjögren's Disease Detection

Bruna Ferreira dos Santos and Lilian Berton

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

brunafdos@gmail.com,lberton@unifesp.br

Abstract. *Sjögren's disease is an autoimmune disorder that primarily affects moisture-producing glands, leading to symptoms such as dry eyes, dry mouth, and systemic complications. Accurate diagnosis remains a challenge, particularly in medical imaging applications. In this study, we employ various Convolutional Neural Networks (CNNs) for Sjögren's disease detection, using a dataset with inherent biases. The goal is to analyze the impact of transfer learning on fairness, investigating whether pre-trained models amplify or mitigate biases in medical image classification. We compare different CNN architectures and evaluate fairness metrics to assess performance disparities across demographic subgroups. The findings contribute to the development of bias-aware AI models, ensuring more equitable and reliable deep learning applications in health.*

1. Introduction

Sjögren's disease is a chronic autoimmune disorder characterized by the body's immune system mistakenly attacking moisture-producing glands, leading to symptoms such as dry eyes, dry mouth, and systemic complications affecting organs like the lungs and nervous system [Nocturne and Mariette 2013]. Early and accurate diagnosis is crucial for managing the disease and preventing severe complications.

Recent advancements in neural networks and deep learning have enabled image-based detection of Sjögren's disease through medical imaging techniques such as salivary gland ultrasonography and histopathological analysis [Álvarez Troncoso et al. 2024, Olivier et al. 2023]. These models utilize convolutional neural networks (CNNs) to identify patterns indicative of the disease, enhancing diagnostic precision and reducing the dependency on traditional, invasive tests.

In the field of image classification, it is common to use pre-trained models as a foundation for various tasks [Schumann et al. 2019]. These models, typically trained on large-scale datasets, offer strong feature extraction capabilities, enabling efficient transfer learning for domain-specific applications. However, despite their advantages, pre-trained models may inadvertently reinforce biases present in the datasets used during training.

Fairness remains a critical challenge in AI-driven diagnosis. Biases in training datasets, such as disproportionate representation of certain demographic groups, can lead to disparities in accuracy, potentially resulting in lower detection rates for underrepresented populations [Rabonato and Berton 2024]. Addressing these concerns requires careful dataset curation, algorithmic adjustments, and fairness-aware evaluation metrics to ensure equitable healthcare outcomes for all patients.

The main goal of this study is to employ various CNNs for Sjögren's disease detection, applied to a dataset with inherent biases. Additionally, this work seeks to assess the impact of transfer learning on fairness, evaluating whether pre-trained models reinforce or mitigate biases in medical image classification.

The paper is organized as follows: Section 2 presents the problem definition. Section 3 the related work. Section 4 describes the methodology. Section 5 presents the results and discussion. Finally, Section 6 presents the conclusions and directions for future work.

2. Problem Definition

When applying transfer learning in medical imaging, especially using convolutional neural networks (CNNs) pre-trained on large-scale natural image datasets like ImageNet, significant distributional mismatches can arise between the source and target domains. Two of the most critical types of domain shift in this context are Subpopulation Shift and Covariate Shift [Guan and Liu 2021].

Subpopulation Shift refers to changes in the composition of subgroups between the source domain (e.g., ImageNet) and the target medical domain. In this scenario, models like VGG, pre-trained on ImageNet, are exposed primarily to non-clinical images that lack sufficient representation of certain demographic or clinical subgroups, such as elderly individuals, people with darker skin tones, or patients with rare medical conditions. As a result, the learned representations may not generalize well to these underrepresented groups in the medical domain. This can lead to two fairness-related concerns: 1) Representational harm, where embeddings for minority subgroups are poorly structured or collapsed in feature space, and 2) Allocation harm, where model predictions disproportionately fail or behave suboptimally for these groups.

A real-world example includes poor performance of VGG-based models when transferred to mammography tasks, particularly in identifying tumors in younger women, a subgroup poorly represented in the original training data.

Covariate Shift, on the other hand, arises when the distribution of input features $P(X)$ changes between domains, even though the conditional relationship $P(Y \mid X)$ remains stable. In this case, features extracted by VGG, which are tuned for natural textures, edges, and lighting conditions, may not be well-suited for the unique characteristics of medical images such as MRIs or ultrasounds. These clinical images often exhibit domain-specific patterns, like grayscale textures, anatomical structures, or imaging artifacts that differ significantly from natural photos. This mismatch in feature representation can lead to reduced model accuracy and unintended biases, especially if the target domain includes subgroups whose data distribution lies far from the source domain's learned manifold. For example, anatomical features common in MRI scans are absent in ImageNet, making VGG's embeddings potentially unreliable in downstream medical tasks.

Together, these domain shifts pose a critical challenge to fairness and generalization in transfer learning for medical AI. Addressing them requires careful analysis, domain adaptation techniques, and fairness-aware evaluation.

3. Related Works

Although several recent studies have explored the application of machine learning in Sjögren's Syndrome (SS), none have specifically addressed potential biases in the datasets or the predictions generated by these models.

3.1. Sjögren's Syndrome by Machine Learning

[Álvarez Troncoso et al. 2024] explores the use of an auto-machine learning (autoML) platform to automatically segment and quantify Focus Score (FS) in histopathological slides of minor salivary gland biopsies, aiming to improve diagnostic accuracy for Sjögren's Syndrome (SS). Using a dataset of 172 slides from 86 patients with sicca symptoms, the model based on ResNet-152, achieved high reliability (score 0.88), sensitivity (89.47%), and specificity (88.24%) in distinguishing SS.

[Wu et al. 2024] introduces CTG-PAM, a graph-based AI model developed to enhance the diagnosis of SS. Traditional diagnostic approaches via histopathology face limitations, prompting the use of 100 whole-slide images from labial gland biopsies to test CTG-PAM. By analyzing features at the single-cell, cell-cell, and cell-tissue levels, the model effectively identifies lymphocytes and performs SS diagnosis using graph-based structure analysis. CTG-PAM outperformed traditional deep learning models like ResNet-50, achieving exceptional diagnostic performance, with an AUC of 1.0 in internal validation and 0.8035 in external testing, and a sensitivity of 98.21%.

[Vyas et al. 2024] explores the use of Raman hyperspectroscopy combined with machine learning to develop a diagnostic tool for SS. It involves analyzing saliva using Raman spectroscopy. These spectral features are then processed through machine learning algorithms and chemometric techniques, which enhance sensitivity and extract disease-specific patterns. They used Genetic Algorithm (GA) for feature selection to reduce dimensionality of Raman spectral data and Support Vector Machine–Discriminant Analysis (SVM-DA) for final spectral classification. Demonstrated 86% sensitivity at the spectral level and 97% accuracy at the sample level.

[Olivier et al. 2023] evaluate a convolutional neural network architecture inspired by U-Net with radiomics-based feature extraction and propose a two-phase deep learning approach, first pretraining with segmentation, then joint classification supervision. This strategy yields segmentation results comparable to human experts and achieves strong diagnostic performance.

3.2. Fairness in transfer Learning

Since models trained in one domain may be used in unexpected settings, raising concerns about whether they still make fair predictions. Some works previously analyzed fairness in transfer learning.

Traditional debiasing methods focus on a single domain, but in real-world applications, especially with pre-trained models or APIs, sensitive attributes may be unavailable, requiring proxy data and alternate fairness evaluations. The authors [Schumann et al. 2019] frame this issue as a domain adaptation problem, proposing a theoretical framework and modeling approach to transfer fairness knowledge from a source domain to a target domain without retraining from scratch. Their method comes with

new fairness guarantees and demonstrates, through empirical results, that fairness across domains can be improved even with limited data.

[Teo et al. 2023] introduces fairTL and fairTL++, two transfer learning approaches designed to mitigate bias in deep generative models. Traditional methods rely on augmenting large, biased datasets with smaller, unbiased reference sets and often require access to all training data. In contrast, fairTL pre-trains a model on a large biased dataset, then adapts it using the smaller, fair dataset, effectively transferring expressive generation capabilities while aligning with fair distributions. fairTL++ enhances this method by incorporating multiple feedback signals and a Linear-Probing followed by Fine-Tuning (LP-FT) strategy. Notably, both approaches remain effective even in constrained scenarios where only a pre-trained model is available and the original dataset is inaccessible, a situation previous methods struggle to address.

4. Materials and Methods

4.1. Dataset

The dataset used in this study consists of 225 medical images related to Sjögren's disease, capturing key biomarkers through techniques such as salivary gland ultrasonography and histopathological analysis. The dataset contains a label that indicates whether the image corresponds to a patient with Sjögren's disease. A label value of 1 denotes a positive case, while 0 indicates a negative case. One image was found to have an invalid label (value = 2) and was therefore removed. As a result, the final dataset consists of 224 images.

Figure 1 illustrates the gender and age distribution across positive and negative diagnoses. It is notable that the proportion of men and individuals under forty years old is relatively small. Figure 2 illustrates the age distribution by gender. The dataset reveals an absence of men in the 50 to 80-year age range, while a few men are present at 20 years old. In contrast, women appear in the dataset starting from 30 years old.

The data was collected from four medical units across Europe, with patients originating from Italy, Serbia, and Slovenia [Zabotti et al. 2020].

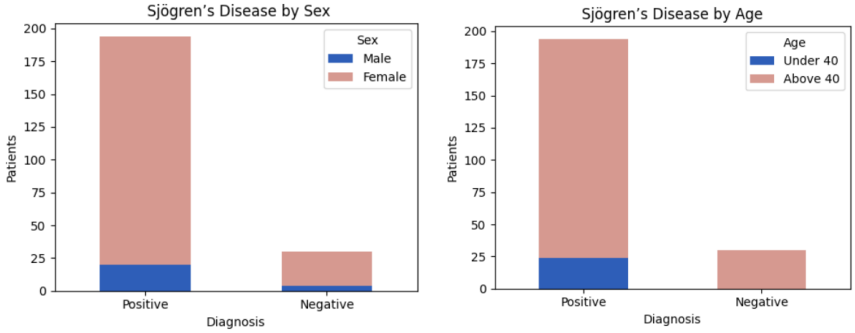


Figure 1. Gender and age distribution among the diagnosis.

4.2. Model Selection and Transfer Learning

Various Convolutional Neural Networks (CNNs) are employed, including pre-trained models on ImageNet (ResNet, EfficientNet, VGG), ResNet pre-treined on RadImageNet

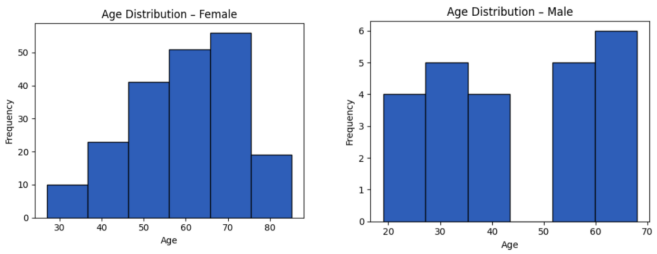


Figure 2. Age distribution by gender.

and custom-built architectures. Transfer learning is applied by leveraging weights pre-trained on ImageNet. The study investigates whether pre-trained models exacerbate or mitigate biases when adapting to a specialized medical dataset.

4.3. Evaluation Metrics

To evaluate bias, the models are assessed using Accuracy, Precision, Recall, F1-score, and AUC for each subgroup, ensuring a comprehensive analysis of performance disparities across different demographic groups.

4.4. Experimental Setup

Three neural network architectures were employed in the experiments: VGG16, ResNet50, and EfficientNet. These models were initialized with pre-trained weights from ImageNet to enable transfer learning. For all architectures, the loss function used was CrossEntropyLoss, and the Adam optimizer was applied to update model parameters, with a learning rate of 1e-4 to control the magnitude of the weight updates.

To evaluate the effectiveness of domain-specific transfer learning, an additional experiment was conducted using the ResNet50 architecture with pretrained weights from RadImageNet. Unlike conventional models pretrained on ImageNet, RadImageNet is trained exclusively on medical images, potentially improving the model's ability to extract clinically relevant features [Mei et al.].

A baseline experiment was also performed using a simple convolutional neural network trained from scratch, without the use of transfer learning.

All models were trained for 50 epochs. Four *seed* values (0, 1, 3, 10, and 42) were tested, and the one that achieved the best metrics was selected. The use of a seed ensures the reproducibility of the experiments, as it maintains the same parameters and conditions when the training is run again. Due to class imbalance in the dataset, the *compute_class_weight* function was used to calculate the weight of each class based on label distribution. These weights were incorporated into the loss function to reduce bias during training and improve the model's generalization ability.

The dataset was split into training and testing subsets, with 70% of the images used for training and 30% for evaluation. The split was performed using the *train_test_split* function with the *stratify* parameter enabled to ensure class balance was maintained across both subsets.

5. Results

This section presents the evaluation metrics for each demographic group analyzed.

5.1. Train Test Split

Table 1 shows how the sensitive variables were splited into training and testing.

		Train	Test
Sex	Female	140	60
	Male	16	8
Age	Under 40	17	7
	Above 40	139	61

Table 1. Distribution of training and testing samples by sex and age group

Figure 3 illustrates the distribution of classes across the training and test sets, while Figure 4 presents the distribution of gender and age within the split.

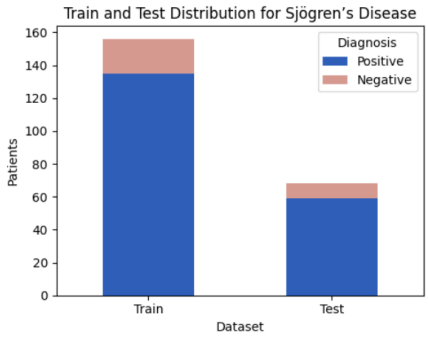


Figure 3. Proportion of labels in train test split.

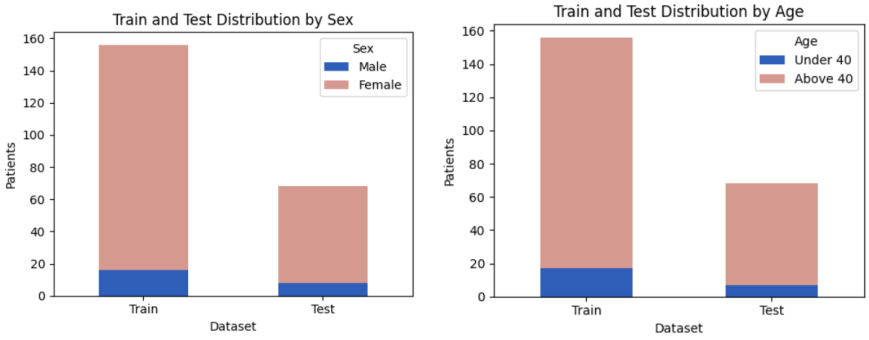


Figure 4. Proportion of gender and age in train test split.

5.2. Classification Results

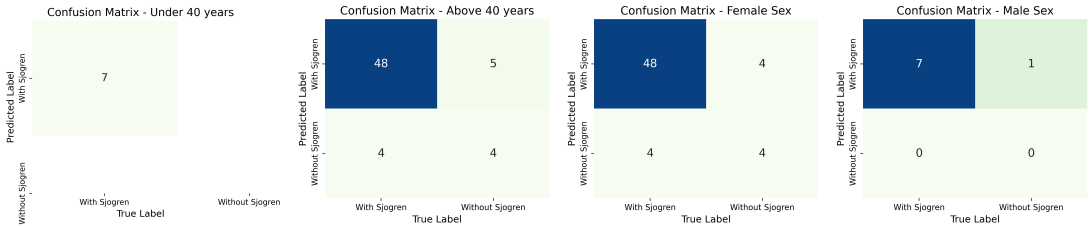
5.2.1. VGG16

The model performs reasonably well on the full dataset (see Table 2 and Figure 5), with high values for F1-score (92.44), indicating good balance between Precision (91.67) and Recall (93.22), suggesting it is effective at detecting positive cases. However, the AUC is relatively low (68.83), indicating limited discriminative ability despite the high classification performance metrics.

For Age $\leq$ 40, all metrics are at 100%, due to all images refers to women. This also explains why the AUC is NaN, likely caused by the absence of positive or negative

examples in the ROC calculation. For Age > 40, the model's performance closely reflects the overall dataset with high F1 (91.43) and Recall (92.31). Slightly lower AUC (68.38), consistent with the full dataset, again reflecting limited probabilistic separation.

Female group shows slightly better AUC (71.15) than the overall dataset and high balanced metrics, indicating robust performance. Male group shows a notably poor AUC (50.00), equivalent to random guessing, despite showing perfect Recall (100%) and good F1 (93.33). This mismatch suggests sample imbalance or low count may be skewing the results. Or the model is overfitting to a specific pattern in male samples, always predicting one class.



(a) Age ≤ 40 **(b) Age > 40** **(c) Female** **(d) Male**

Figure 5. Confusion matrices for age and sex subgroups in the test set for VGG16.

Metric	Full Dataset	Age ≤ 40 years	Age > 40 years	Female	Male
AUC	68.83	NaN	68.38	71.15	50.00
F1	92.44	100.00	91.43	92.31	93.33
Accuracy	86.76	100.00	85.25	86.67	87.50
Precision	91.67	100.00	90.57	92.31	87.50
Recall	93.22	100.00	92.31	92.31	100.00

Table 2. Test Performance Metrics - VGG16

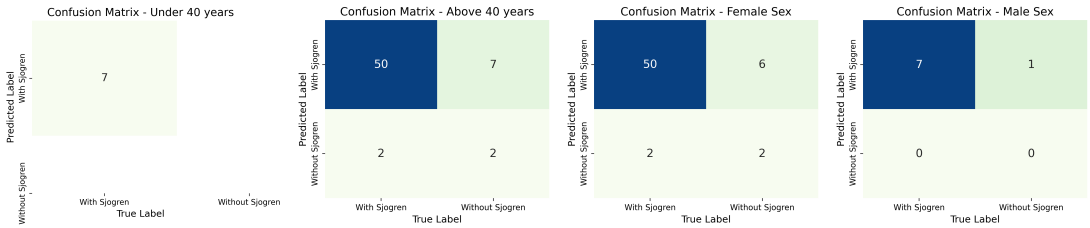
5.2.2. Resnet50

The model shows high classification metrics on the full dataset (see Table 3 and Figure 6). F1-score of 92.68, Recall of 96.61, and Accuracy of 86.76%, indicating strong performance in identifying positive cases. However, the AUC is 59.42, which is low and suggests that, despite high accuracy and recall, the model struggles to separate classes probabilistically.

For the Age ≤ 40 group, all metrics are 100%, but the AUC is NaN, likely due to no men under 40. For Age > 40, the metrics (F1-score: 91.74, Recall: 96.15, Accuracy: 85.25) are consistent with the full dataset. Yet again, the AUC remains low (59.19), reinforcing the concern about class separability in probabilistic terms.

For the female group, the model maintains strong F1 (92.59), Recall (96.15), and Accuracy (86.67), along with a slightly improved AUC of 60.58, though still underwhelming. For the male group, the model achieves perfect Recall (100%), strong F1 (93.33), and Accuracy (87.50). However, the AUC is 50.00, indicating no probabilistic distinction between classes. This mismatch suggests the model may be overconfidently predicting one class for this subgroup.

The high Recall and F1-scores across all groups show the model can correctly identify positive cases. However, low AUC values across the board, particularly 50.00 for males and NaN for ≤ 40, indicate sample size imbalance.



(a) Age ≤ 40 **(b) Age > 40** **(c) Female** **(d) Male**

Figure 6. Confusion matrices for age and sex subgroups in the test set for ResNet.

Metric	Full Dataset	Age ≤ 40 years	Age > 40 years	Female	Male
AUC	59.42	NaN	59.19	60.58	50.00
F1	92.68	100.00	91.74	92.59	93.33
Accuracy	86.76	100.00	85.25	86.67	87.50
Precision	89.06	100.00	87.72	89.29	87.50
Recall	96.61	100.00	96.15	96.15	100.00

Table 3. Performance metrics by subgroup for ResNet50 model

5.2.3. EfficientNet

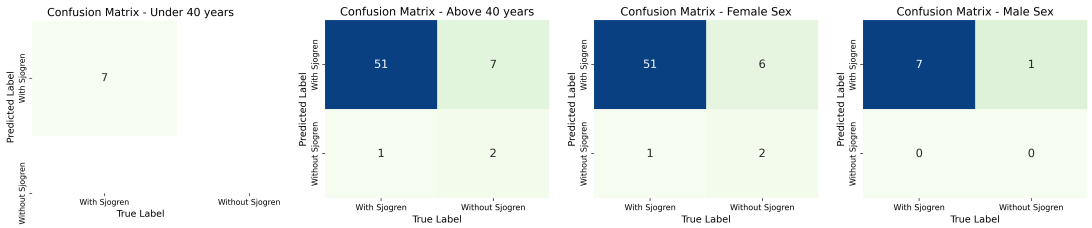
As shown in Table 4 and Figure 7, F1-score is high at 98.31, indicating a strong balance between precision and recall. Accuracy is also solid at 88.24%, confirming the model's strong general predictive performance. Recall (98.31) is slightly higher than Precision (89.23), suggesting the model favors correctly identifying positive instances, potentially at the cost of more false positives. However, AUC is only 60.26, which is quite low and suggests the model has poor probabilistic separability between classes.

For the Age ≤ 40 group all metrics are 100%, including F1, Precision, and Recall. However, AUC is NaN, indicating that the ROC curve could not be computed, likely due to no man are in this group. For the Age > 40 group, F1-Score(92.73), Precision (87.93), and Recall (98.08) remain strong and closely mirror the full dataset. AUC is still low at 60.15, reinforcing that the model struggles to rank predictions by confidence.

For female subgroup, it achieve very good metrics: F1-Score(93.58), Accuracy (88.33), Recall (98.08), and Precision (89.47). AUC is slightly higher than average at 61.54, but still well below ideal. For male subgroup it achieved perfect Recall (100%) and strong F1-Score(93.33) and Accuracy (87.50). However, AUC drops to 50.00, suggesting the model is no better than random at ranking predictions for males, likely a result of overconfident or biased outputs.

EfficientNet achieves excellent classification metrics (F1, Recall, Accuracy) across all groups, suggesting it learns patterns well. Low AUC values (especially 50.00 for males and NaN for younger age group) indicate poor probabilistic calibration, which

may result in unreliable confidence estimates. The perfect scores in small subgroups hint at potential overfitting or data imbalance.



(a) Age ≤ 40 (b) Age > 40 (c) Female (d) Male

Figure 7. Confusion matrices for age and sex subgroups in the test set for EfficientNet.

Metric	Full Dataset	Age ≤ 40 years	Age > 40 years	Female	Male
AUC	60.26	NaN	60.15	61.54	50.00
F1-Score	93.55	100.00	92.73	93.58	93.33
Accuracy	88.24	100.00	86.89	88.33	87.50
Precision	89.23	100.00	87.93	89.47	87.50
Recall	98.31	100.00	98.08	98.08	100.00

Table 4. Performance metrics by subgroup for EfficientNet model

5.2.4. RadImageNet - Resnet50

The distinguishing feature of this subsection is the use of RadImageNet, a pre-trained dataset specifically designed for the medical field. It is analogous to ImageNet but focused on medical images (e.g., CT scans, X-rays, MRIs). Using pre-trained weights from RadImageNet represents a form of specialized transfer learning, where the model begins training with prior knowledge tailored to the radiology domain. This approach aims to offer better adaptation to the medical domain compared to traditional ImageNet pre-training.

The performance metrics for the RadImageNet model, shown in Table 5 and Figure 8), reveal some notable subgroup differences. The overall AUC for the full dataset is moderate at 63.28, with subgroup AUCs ranging from 50.00 for males to 64.90 for females, indicating a slight gender disparity in discriminative ability. Interestingly, the AUC for the age group 40 years is not available (NaN), which suggests either insufficient data or an issue with metric calculation in this subgroup.

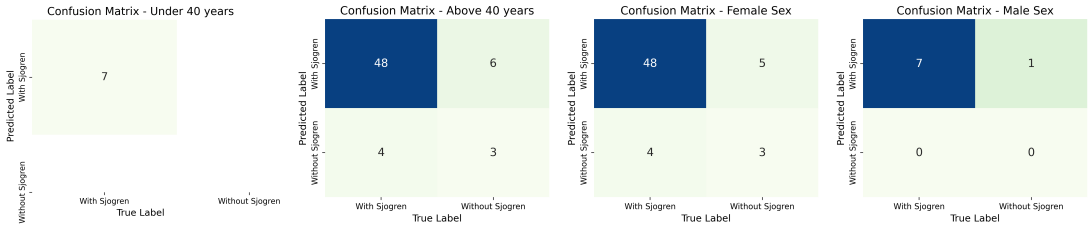
In contrast, the other classification metrics, F1, accuracy, precision, and recall are quite high across most groups. Notably, the subgroup aged ≤ 40 years achieves perfect scores (100%) on F1, accuracy, precision, and recall, likely due to have only women in this group. Meanwhile, males show the highest recall (100%) but slightly lower AUC and precision compared to females, implying that while the model captures all positive cases in males, it may produce more false positives.

When comparing the results of the RadImageNet model with those of the EfficientNet model, it is observed that both present relatively similar performance. However, EfficientNet demonstrates a slight superiority in most metrics.

In the full dataset, EfficientNet achieved an AUC of 60.26, slightly lower than RadImageNet's 63.28. This pattern is also observed in the subgroups of age ¿ 40 years and females, where RadImageNet presents marginally higher AUCs (60.15 vs. 62.82 and 61.54 vs. 64.90, respectively). In both models, the AUC for the male group remains at 50.00, indicating low discriminative ability for this group in both cases.

Regarding the other metrics, EfficientNet outperforms RadImageNet. The F1-Score on the test dataset is 93.55 for EfficientNet, compared to 91.67 for RadImageNet. This difference is consistent in the subgroups of age ¿ 40 years and females (92.73 vs. 90.57 and 93.58 vs. 91.43, respectively).

RadImageNet presents competitive performance but does not lead in any metric or subgroup. VGG16 stands out particularly in AUC, EfficientNet in F1 and Recall, while ResNet50 maintains results close to RadImageNet but with a slight advantage in some scenarios. For the male group, all models perform equally poorly (AUC = 50.00).



(a) Age $\leq$ 40 (b) Age $>$ 40 (c) Female (d) Male

Figure 8. Confusion matrices for age and sex subgroups in the test set using the RadImageNet model.

Metric	Full Dataset	Age $\leq$ 40 years	Age $>$ 40 years	Female	Male
AUC	63.28	NaN	62.82	64.90	50.00
F1-Score	91.67	100.00	90.57	91.43	93.33
Accuracy	85.29	100.00	83.61	85.00	87.50
Precision	90.16	100.00	88.89	90.57	87.50
Recall	93.22	100.00	92.31	92.31	100.00

Table 5. Performance metrics by subgroup for RadImageNet model

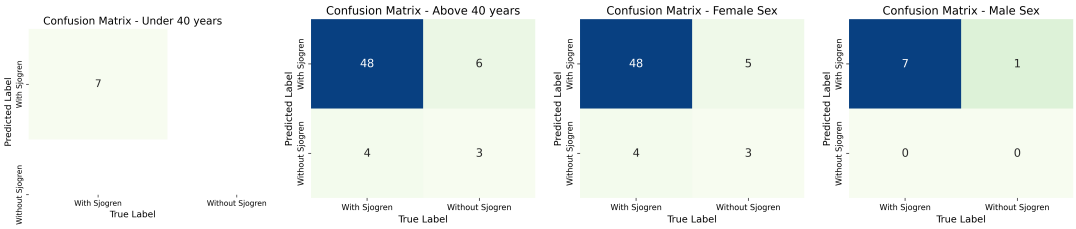
5.2.5. CNN

A comparative analysis between the CNN model and the other evaluated models shows that, in terms of AUC, the CNN exhibits performance equivalent to RadImageNet in all scenarios, indicating virtually identical behavior between the two. In the full dataset, the CNN achieves an AUC of 63.28, outperforming ResNet50 (59.42) and EfficientNet (60.26), but falling short of VGG16 (68.83). The same pattern is observed in the subgroups of individuals over 40 years of age and females, with the CNN obtaining values of 62.82 and 64.90, respectively, which are higher than those of ResNet50 and EfficientNet, yet lower than those of VGG16. For the male group, all models show an AUC of 50.00, highlighting low discriminative ability for this subgroup.

Regarding the F1-Score, the CNN matches RadImageNet's performance but remains below that of VGG16, ResNet50, and EfficientNet. In the full dataset, the CNN

reaches 91.67, while EfficientNet achieves 93.55 and VGG16 scores 92.44. This pattern is consistent in the subgroups of individuals over 40 years and females, where EfficientNet maintains the highest values, followed by ResNet50 and VGG16. In terms of accuracy, the CNN records the same results as RadImageNet (85.29 in the full dataset), which are lower than those of VGG16 and ResNet50 (86.76) and EfficientNet (88.24).

For precision and recall, the CNN also mirrors RadImageNet's performance. In precision, it achieves 90.16 in the full dataset, slightly outperforming ResNet50 (89.06) and EfficientNet (89.23), but remaining below VGG16 (91.67). For recall, the CNN records 93.22 in the full dataset, equal to VGG16, but lower than ResNet50 (96.61) and, most notably, EfficientNet (98.31). These results indicate that although the CNN and RadImageNet are competitive compared to the other models, they do not lead in any key metric, being surpassed by VGG16 in AUC and by EfficientNet in F1-Score and recall.



(a) Age $\leq$ 40 (b) Age $>$ 40 (c) Female (d) Male

Figure 9. Confusion matrices for age and sex subgroups in the test set using the CNN model.

Metric	Full Dataset	Age $\leq$ 40 years	Age $>$ 40 years	Female	Male
AUC	63.28	NaN	62.82	64.90	50.00
F1-Score	91.67	100.00	90.57	91.43	93.33
Accuracy	85.29	100.00	83.61	85.00	87.50
Precision	90.16	100.00	88.89	90.57	87.50
Recall	93.22	100.00	92.31	92.31	100.00

Table 6. Performance metrics by subgroup for the CNN model.

6. Conclusion

The evaluation of multiple models, VGG16, ResNet50, EfficientNet, RadImageNet (ResNet-based), and CNN, reveals that pre-trained models, including those based on general (ImageNet) or domain-specific (RadImageNet) datasets, do not consistently reduce bias across demographic subgroups. While these models achieve high classification metrics (F1, recall, accuracy) overall, they suffer from low AUC scores, especially for the male subgroup, indicating poor probabilistic calibration and limited discriminative ability. While pretraining, especially with RadImageNet, can offer slight improvements in alignment with the medical domain, it does not inherently mitigate subgroup bias. Performance disparities across age and gender highlight the importance of addressing data representativeness and fairness explicitly, rather than relying solely on transfer learning to resolve them.

7. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2021/14725-3.

References

Álvarez Troncoso, J., Ruiz-Bravo, E., Soto Abánades, C., Dumusc, A., López-Janeiro, Á., and Hügle, T. (2024). Classification of salivary gland biopsies in sjögren's syndrome by a convolutional neural network using an auto-machine learning platform. *BMC rheumatology*, 8(1):60.

Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185.

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z. A., and Yang, Y. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315.

Nocturne, G. and Mariette, X. (2013). Advances in understanding the pathogenesis of primary sjögren's syndrome. *Nature Reviews Rheumatology*, 9(9):544–556.

Olivier, A., Hoffmann, C., Jousse-Joulin, S., Mansour, A., Bressollette, L., and Clement, B. (2023). Machine and deep learning approaches applied to classify gougerot–sjögren syndrome and jointly segment salivary glands. *Bioengineering*, 10(11):1283.

Rabonato, R. T. and Berton, L. (2024). A systematic review of fairness in machine learning. *AI and Ethics*, pages 1–12.

Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., and Chi, E. H. (2019). Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*.

Teo, C. T., Abdollahzadeh, M., and Cheung, N.-M. (2023). Fair generative models via transfer learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2429–2437.

Vyas, B., Khatiashvili, A., Galati, L., Ngo, K., Gildener-Leapman, N., Larsen, M., and Lednev, I. K. (2024). Raman hyperspectroscopy of saliva and machine learning for sjögren's disease diagnostics. *Scientific Reports*, 14(1):11135.

Wu, R., Chen, Z., Yu, J., Lai, P., Chen, X., Han, A., Xu, M., Fan, Z., Cheng, B., Jiang, Y., et al. (2024). A graph-learning based model for automatic diagnosis of sjögren's syndrome on digital pathological images: a multicentre cohort study. *Journal of Translational Medicine*, 22(1):748.

Zabotti, A., Callegher, S., Tullio, A., Vukicevic, A., Hocevar, A., Milic, V., Cafaro, G., Carotti, M., Delli, K., De Lucia, O., Ernst, D., Ferro, F., Gattamelata, A., Germanò, G., Giovannini, I., Hammenfors, D., Jonsson, M., Jousse-Joulin, S., Macchioni, P., and Vita, S. (2020). Salivary gland ultrasonography in sjögren's syndrome: A european multicenter reliability exercise for the harmonicss project. *Frontiers in Medicine*, 7:581248.