This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Towards Privacy-Preserving Multi-Site ICH Detection for South African Telestroke Networks

Jaime Leigh Kruger

Stellenbosch University, Stellenbosch, South Africa
24814415@sun.ac.za

Abstract. Intracranial haemorrhage (ICH) is a life-threatening neurological emergency requiring rapid identification on non-contrast computed tomography (CT). In South Africa, where radiologist scarcity limits timely interpretation, deep learning offers automated detection to support clinical triage. However, deploying such models across multiple hospitals faces two challenges: domain shift from heterogeneous CT scanners degrades accuracy, and the Protection of Personal Information Act (POPIA) prohibits centralised data pooling. We present a framework addressing both challenges through unsupervised domain adaptation and federated learning (FL). Using a DenseNet-121 classifier, we evaluated five adaptation methods and three FL training strategies on public proxy datasets. Deep CORAL was the only adaptation method that did not degrade cross-site performance, showing a directionally positive effect (+1.0 pp, though not statistically significant at $n = 322$) without requiring target labels. For multi-site training, Federated Averaging (FedAvg) matched centralised performance under realistic non-IID conditions. Adding differential privacy via DP-SGD provided formal guarantees (ε from 4.59 to 0.13) at a substantial accuracy cost that our results attribute largely to the required architectural change rather than the privacy noise. These results suggest a feasible pathway for privacy-preserving ICH detection in resource-constrained settings.

Keywords: Medical AI for Africa · Domain adaptation · Federated learning · Intracranial haemorrhage · Telestroke

1 Introduction

Stroke is the leading cause of adult disability in sub-Saharan Africa, with haemorrhagic subtypes accounting for approximately 34% of cases — substantially higher than the 15–20% observed in high-income countries [4,6]. Rapid identification of intracranial haemorrhage (ICH) on non-contrast computed tomography (CT) is critical: tissue plasminogen activator (tPA) administration is absolutely contraindicated in haemorrhagic stroke, and misclassification can be fatal.

Currently, ICH detection at district hospitals depends on radiologist interpretation. However, sub-Saharan Africa has approximately 1 radiologist per 100,000

population compared to 12 per 100,000 in the United States [15], resulting in delayed diagnoses and suboptimal triage at peripheral facilities. Deep learning models have demonstrated radiologist-level accuracy for ICH detection on large benchmark datasets [3], offering a potential solution to this workforce gap. However, deploying such models across South African hospitals faces two structural barriers.

Domain shift. Models trained on international datasets degrade when applied to local scanner hardware due to differences in acquisition parameters, reconstruction kernels, and noise characteristics — a well-documented problem in medical imaging deployment [21]. The radiologist scarcity that motivates automation simultaneously makes manual annotation of local training data infeasible.

Data privacy. The Protection of Personal Information Act (POPIA) prohibits sharing identifiable patient data between institutions without explicit consent, legally constraining centralised multi-site training.

We present a combined framework: (1) **Domain adaptation** via Deep CORAL [17] to close the cross-scanner gap using only unlabelled scans; (2) **Federated learning** via FedAvg [14] to enable multi-site collaboration without data centralisation; (3) **Differential privacy** via DP-SGD [1] to provide formal privacy guarantees. We evaluated on public proxy datasets (RSNA→CQ500) and demonstrated feasibility for resource-constrained deployment.

2 Clinical Context

South Africa faces high stroke incidence with younger onset (mean 55 vs. 73 in Europe), driven by hypertension and HIV-associated vasculopathy [4]. District hospitals provide primary CT access but typically operate a single scanner without on-site specialists. The deployment environment imposes specific constraints: (1) scanner heterogeneity across GE, Siemens, Philips, and Canon hardware [20]; (2) no centralised labels due to radiologist scarcity; (3) POPIA data residency requirements; (4) variable connectivity; (5) heterogeneous case mix across referral patterns.

3 Related Work

ICH detection with deep learning. Chilamkurthy et al. [3] demonstrated automated ICH detection achieving AUROC of 0.94 on 313,318 head CT scans. The RSNA 2019 Brain CT Hemorrhage Challenge [7] established a large-scale public benchmark (752,803 slices, 6 subtypes), on which DenseNet-121 [10] achieves $>$0.97 AUROC. However, these results assume matched training and deployment distributions.

Domain adaptation. Domain adaptation transfers knowledge from a labelled source to an unlabelled target domain. Ben-David et al. [2] bound target error by source error plus domain divergence. For medical imaging, shift arises from scanner hardware, protocols, and demographics. Deep CORAL [17] aligns second-order feature statistics without target labels. Domain-Adversarial Neural Networks (DANN) [8] use gradient reversal to learn domain-invariant features. Source Hypothesis Transfer (SHOT) [13] adapts without source data access. Test-time methods including TENT [18] and CoTTA [19] operate at inference with no retraining. No prior work has systematically compared these methods for cross-scanner ICH detection.

Federated learning in healthcare. Federated learning (FL) enables collaborative training without centralising data [14,16]. The EXAM study [5] demonstrated FL across 20 institutions for COVID-19 prognosis, but operated within well-resourced US networks. Healthcare federations in low- and middle-income countries (LMICs) face greater heterogeneity: Hsu et al. [9] showed Dirichlet label skew ($\alpha < 1.0$) degrades FedAvg convergence. FedProx [11] adds proximal regularisation; FedBN [12] keeps batch normalisation local. DP-SGD [1] provides formal per-sample privacy guarantees, though at a utility cost poorly characterised for medical imaging.

4 Methods

Our framework comprised three components: (1) a baseline ICH classifier; (2) unsupervised domain adaptation for cross-scanner deployment; and (3) federated learning with optional differential privacy for multi-site collaboration.

4.1 Baseline Model

We used DenseNet-121 [10] with ImageNet-pretrained weights, replacing the final fully connected layer with a 6-unit sigmoid output (one per ICH subtype). Input preprocessing converted each DICOM slice into a 3-channel image using CT windows: brain (W:80/L:40), subdural (W:200/L:80), and bone (W:2800/L:600), resized to 224×224. Training used Adam (10^{-4}), batch size 32 (maximising V100 16 GB memory), and binary cross-entropy for 5 epochs — sufficient because ImageNet pre-training provided strong initial features and validation AUROC plateaued after epoch 3. This baseline (standard DenseNet-121, no architectural modifications) achieved 0.975 AUROC on the RSNA 2019 test set [7] (752,803 slices, 13,256 training patients).

4.2 Domain Adaptation via Deep CORAL

Deep CORAL [17] minimises the Frobenius norm between source and target feature covariance matrices (results in Table 1):

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(x_s, y_s) + \lambda \cdot \frac{1}{4d^2} \|C_s - C_t\|_F^2 \tag{1}$$

with $\lambda = 1.0$ (Sun & Saenko's original default, not tuned on CQ500) and $d = 1024$. CORAL was selected from a systematic comparison of five methods (CORAL, DANN, SHOT, TENT, CoTTA) as the only one that did not degrade performance on this task. All methods used their respective papers' default hyperparameters: DANN used the standard gradient reversal schedule of Ganin et al. [8] with $\gamma = 10$; SHOT, TENT, and CoTTA followed [13,18,19]. All methods used learning rate 10^{-4} (source-dependent) or 10^{-5} (test-time), 7 epochs for source-dependent and 1–3 epochs for test-time methods.

4.3 Federated Learning

FedAvg [14]: each client trained locally for 1 epoch per round, uploaded parameters to a coordinator, and received the weighted average:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)} \tag{2}$$

We partitioned $K = 5$ clients using Dirichlet(α) sampling [9] over the binary ICH label. We also evaluated FedProx [11] ($\mu = 0.01$) and FedBN [12].

4.4 Differential Privacy

DP-SGD [1] provided formal per-sample privacy guarantees via gradient clipping (max norm $C = 1.0$) and calibrated Gaussian noise ($\sigma \in \{0.5, 1.0, 2.0\}$). BatchNorm layers were replaced with GroupNorm (required for per-sample gradients). Privacy budget was computed via Rényi differential privacy accounting ($\delta = 10^{-5}$).

5 Experiments and Results

5.1 Domain Adaptation: RSNA → CQ500

We evaluated transfer from RSNA 2019 (US) to CQ500 [3] (India, 322 studies, 134 ICH-positive). The proxy A-distance between domains is 1.90 (computed as $2(1 - 2\epsilon_{\text{disc}})$ where ϵ_{disc} is the error of a linear SVM on penultimate-layer features [2]), confirming substantial feature-level shift (Figure 1). Table 1 presents study-level results.

Deep CORAL was the only method that did not degrade performance, achieving 0.842 AUROC (+1.0 pp) without target labels. The improvement was directionally consistent across evaluation epochs (epochs 4–6 all exceed baseline); bootstrap CIs are wide at this sample size (CORAL: [0.793, 0.889]; baseline: [0.782, 0.881]). DANN caused negative transfer by destroying class-discriminative features overlapping with domain-discriminative ones. SHOT degraded due to unreliable pseudo-labels on the small imbalanced target set. TENT and CoTTA produced no change because the shift resided in convolutional features rather than batch normalisation statistics.

Table 1: Domain adaptation results (DenseNet-121 → CQ500, study-level AUROC, $n = 322$). All methods initialised from the RSNA-trained checkpoint.

Method	Type	CQ500 AUROC	Δ
Baseline (no adaptation)	—	0.832	—
Deep CORAL	Unsupervised	**0.842**	+1.0 pp
DANN	Adversarial	0.822	−1.0 pp
SHOT	Source-free	0.795	−3.7 pp
TENT	Test-time	0.832	0.0 pp
CoTTA	Continual TTA	0.832	0.0 pp

Table 2: Fine-tuning data efficiency vs. CORAL (zero labels required).

Method	Labelled Studies	CQ500 AUROC
No adaptation	0	0.832
Fine-tune (N=10)	10	0.742
Fine-tune (N=50)	50	0.767
Fine-tune (N=100)	100	0.795
Deep CORAL	**0**	**0.842**

Why not supervised fine-tuning? Table 2 quantifies the annotation burden that unsupervised adaptation avoided. Fine-tuning ($\text{lr} = 10^{-5}$, batch 16, 20 epochs, best checkpoint by validation AUROC) with up to 100 labelled studies failed to match the unadapted baseline, while CORAL surpassed all supervised configurations without a single label. The labelled fine-tuning studies were held out of the CQ500 evaluation via a stratified split, so the fine-tuning AUROCs in Table 2 are measured on unseen studies (the held-out remainder) with no fine-tuning data leaking into evaluation.

Figure 1 confirmed that RSNA and CQ500 formed distinct clusters in feature space despite both containing head CT slices. The shift resided in learned representations rather than raw pixel intensities (Jensen–Shannon divergence of intensity histograms: $\text{JSD} = 0.010$).

Figure 2 confirmed that predictions were based on clinically relevant features (haemorrhage regions) rather than scanner artefacts, supporting clinical trust in the baseline model prior to domain adaptation.

5.2 Federated Learning Under Non-IID Conditions

Table 3 presents the federated results under realistic heterogeneity.

FedAvg achieved 0.9776 AUROC, marginally above the centralised model's bootstrap CI [0.974, 0.976] (+0.3 pp over the 0.975 point estimate, a single-run difference), under conditions where client prevalence ranged from 9.7% to 96.6%. FedProx performed within 0.0003 of FedAvg, suggesting the proximal term was

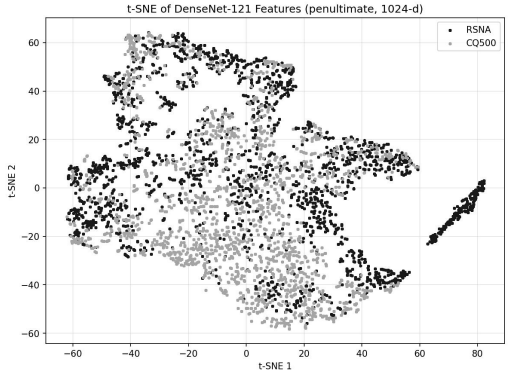


Fig. 1: t-SNE visualisation of DenseNet-121 penultimate-layer features (1024-dimensional) for RSNA and CQ500 samples. Clear domain clustering confirms feature-level shift ($d_{\mathcal{A}} = 1.90$) that CORAL's covariance alignment addresses.

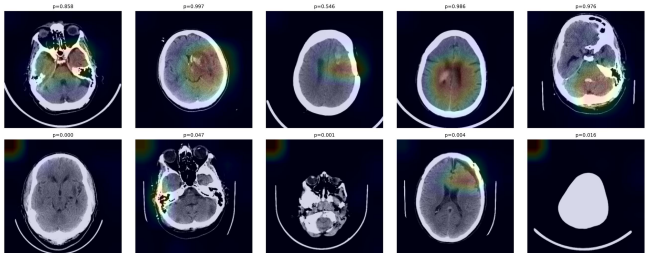


Fig. 2: Grad-CAM heatmaps (baseline DenseNet-121, final convolutional block) on RSNA CT slices. Top: ICH-positive cases ($p > 0.5$) with attention on haemorrhage regions. Bottom: true-negatives ($p < 0.05$) with diffuse activation.

unnecessary when initial model quality was high. FedBN degraded monotonically from 0.9741 (round 1) to 0.9645 (round 5), as local BN statistics accumulated distribution mismatch under label skew.

Heterogeneity robustness. Table 4 shows FedAvg performance across three Dirichlet concentration levels.

FedAvg retained >0.97 AUROC across all tested heterogeneity levels, including extreme $\alpha = 0.1$ where one client received zero samples (excluded, yielding 4 effective clients). Figure 3(a) illustrates the non-IID label distribution at $\alpha = 0.5$, simulating variation across South African hospital types. Figure 3(b) shows convergence: all strategies peaked within 3–4 rounds, with FedBN degrading from round 1.

Table 3: Federated learning aggregation strategies ($\alpha = 0.5$, prevalence 9.7%–96.6%, 5 clients). Centralised baseline: 0.975 [0.974, 0.976]. Single-simulation results.

Strategy	Best AUROC	Best Round	Rounds Done	Δ_{central}
FedAvg	**0.9776**	3–4	5	+0.3 pp
FedProx (μ=0.01)	0.9773	4	5	+0.2 pp
FedBN	0.9741	1	5	−0.1 pp

Table 4: FedAvg robustness across heterogeneity levels. All results lie within 0.3 pp of the centralised point estimate (0.975).

α	Prevalence Range	Best AUROC	Δ_{central}
0.1 (extreme, K_{eff}=4)	0.0%–78.7%	0.9739	−0.1 pp
0.5 (realistic)	9.7%–96.6%	0.9776	+0.3 pp
1.0 (mild)	29.9%–94.0%	0.9779	+0.3 pp

5.3 Differential Privacy: Privacy–Utility Frontier

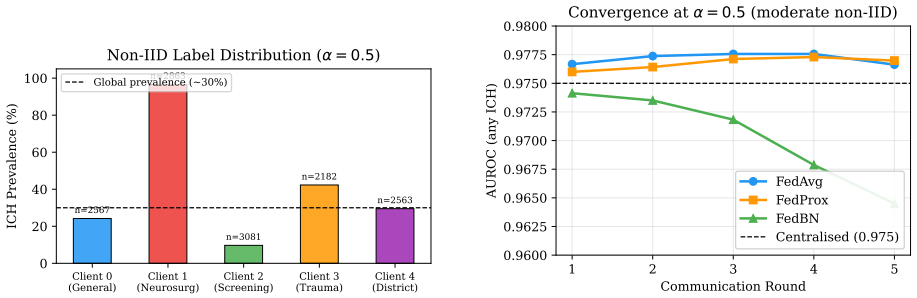
Table 5 presents the privacy–utility trade-off observed when adding formal DP guarantees to federated training.

Formal DP reduced AUROC by approximately 20 pp relative to non-private FedAvg. Increasing the noise multiplier four-fold (σ: 0.5→2.0) cost only 1.5 pp additional AUROC despite reducing ε from 4.59 to 0.13 (35× tighter guarantee), suggesting the dominant cost was the BatchNorm→GroupNorm architectural change rather than noise itself. All DP configurations remained above chance but fell below clinically acceptable thresholds.

6 Discussion

Addressing the African AI gap. Our framework required no local annotation (CORAL used unlabelled data), no centralised infrastructure (federation kept data local), and minimal bandwidth (32 MB per round). Unlike the EXAM study [5], which demonstrated FL across 20 well-resourced US institutions, our setting targeted the constraints of LMIC deployment: fewer labels, greater heterogeneity, and stricter privacy regulation under POPIA. No patient data left institutional boundaries at any stage.

Method selection. When domain shift is scanner-driven rather than demographic, second-order statistical alignment (CORAL) outperformed adversarial (DANN: negative transfer), source-free (SHOT: pseudo-label collapse), and test-time approaches (TENT/CoTTA: wrong layer of shift). FedAvg and FedProx performed indistinguishably (0.9776 vs 0.9773), suggesting strong initialisation reduced sensitivity to the aggregation strategy.



(a) Non-IID label distribution ($\alpha = 0.5$). (b) Convergence within 3–4 rounds.

Fig. 3: Federated learning under realistic heterogeneity simulating South African hospital variation.

Table 5: Privacy–utility frontier for DP-SGD federated training ($\alpha = 0.5$, 5 clients, 5 rounds, $\delta = 10^{-5}$). ε computed via RDP accounting for the worst-case (smallest) client. Non-private baseline uses BatchNorm; DP runs use GroupNorm + gradient clipping.

Configuration	AUROC	ε	$\Delta_{\text{non-private}}$
FedAvg (non-private, BN)	0.9776	∞	—
DP-SGD $\sigma = 0.5$ (GN)	0.7813	4.59	−0.196
DP-SGD $\sigma = 1.0$ (GN)	0.7736	0.64	−0.204
DP-SGD $\sigma = 2.0$ (GN)	0.7666	0.13	−0.211

Privacy–utility trade-off. The 20 pp DP-SGD cost appeared dominated by the GroupNorm architectural change rather than noise — quadrupling σ cost only 1.5 pp additional AUROC despite tightening ε by 35×. Sites can deploy the non-private federated model (already POPIA-compliant) and add DP selectively.

Limitations and future work. We used proxy datasets (RSNA, CQ500) rather than South African data, so the deployment-feasibility claim remains untested in the target setting; a real South African pilot dataset is the priority next step. Federated experiments simulated label-heterogeneity on a single partitioned dataset, testing aggregation under label skew rather than true multi-site federation, which would also introduce feature shift and communication failures. The CORAL improvement (+1.0 pp) was not statistically significant at $n = 322$, and all federated and DP results are single-run; multi-seed evaluation with formal significance testing and metrics beyond AUROC (sensitivity, specificity, F1, precision–recall AUC, and calibration) is left to future work. Isolating the DP accuracy cost between the GroupNorm change and the privacy noise would require a GroupNorm-without-DP baseline. Code is available at `https://github.com/jaimel4/africai2026-ich-detection`.

7 Conclusion

We presented evidence that privacy-preserving multi-site ICH detection is feasible for South African telestroke. Of five adaptation methods, only CORAL did not degrade cross-scanner performance (+1.0 pp, not statistically significant at $n = 322$). FedAvg matched centralised training under realistic non-IID conditions. Formal DP via DP-SGD cost ∼20 pp AUROC, which our results attribute largely to the required architectural change rather than the privacy noise, though isolating this would require a GroupNorm-only baseline. Together, these techniques enable staged deployment requiring no local labels, no data sharing, and minimal compute. Validation on South African hospital data is the necessary next step.

Acknowledgments. Computations were performed using facilities provided by the Centre for High Performance Computing (CHPC), South Africa. The author thanks Dr Mkhuseli Ngxande for supervision.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: ACM CCS. pp. 308–318 (2016). https://doi.org/10.1145/2976749.2978318
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79**(1–2), 151–175 (2010). https://doi.org/10.1007/s10994-009-5152-4
3. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018). https://doi.org/10.1016/S0140-6736(18)31645-3
4. Connor, M.D., Walker, R., Modi, G., Warlow, C.P.: Burden of stroke in black populations in sub-Saharan Africa. The Lancet Neurology **6**(3), 269–278 (2007). https://doi.org/10.1016/S1474-4422(07)70002-9
5. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al.: Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine **27**, 1735–1743 (2021). https://doi.org/10.1038/s41591-021-01506-3
6. Feigin, V.L., Stark, B.A., Johnson, C.O., Roth, G.A., Bisignano, C., Abady, G.G., et al.: Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology **20**(10), 795–820 (2021). https://doi.org/10.1016/S1474-4422(21)00252-0
7. Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P.: Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge. Radiology: Artificial Intelligence **2**(3), e190211 (2020). https://doi.org/10.1148/ryai.2020190211

8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. In: JMLR. vol. 17, pp. 1–35 (2016)
9. Hsu, T.M.H., Qi, H., Brown, M.: Measuring the effects of non-identical data distribution for federated visual classification. In: NeurIPS Workshop on Federated Learning (2019)
10. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
11. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys. pp. 429–450 (2020)
12. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: Federated learning on non-IID features via local batch normalization. In: ICLR (2021)
13. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML. pp. 6028–6039 (2020)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. pp. 1273–1282 (2017)
15. Ogbole, G.I., Adeyomoye, A.O., Bode, C.O., Lands, Y., Ogunseyinde, A.O., Ogunleye, O.T.: Radiology workforce in sub-Saharan Africa: a brief review. World Journal of Radiology **10**(8), 89–96 (2018). https://doi.org/10.4329/wjr.v10.i8.89
16. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ Digital Medicine **3**, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1
17. Sun, B., Saenko, K.: Deep CORAL: Correlation alignment for deep domain adaptation. In: ECCV Workshops. pp. 443–450 (2016). https://doi.org/10.1007/978-3-319-49409-8_35
18. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
19. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022). https://doi.org/10.1109/CVPR52688.2022.00706
20. van Wijk, M., Barnard, M.M., Fernandez, A., Cloete, K., Mukosi, M., Pitcher, R.D.: Trends in public sector radiological usage in the Western Cape Province, South Africa: 2009–2019. South African Journal of Radiology **25**(1), a2251 (2021). https://doi.org/10.4102/sajr.v25i1.2251
21. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-site study using CheXNet. PLOS Medicine **15**(11), e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683