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.

Hierarchical Task Framing for Prevention-Oriented Cervical Lesion Detection Using Nigerian Colposcopy Images

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Cervical cancer progresses through a precancerous stage that, if detected early, can be treated before invasive carcinoma develops. However, most deep learning approaches to colposcopy-based screening frame the problem as binary: normal-versus-abnormal or cancer-versus-non-cancer classification, overlooking the precancerous class that is most actionable for prevention. Whether task framing affects the accurate detection of precancerous lesions remains underexplored. We present a hierarchical task framing for cervical lesion grading that decomposes the problem into a cascade of clinically motivated binary decisions. We compare this against conventional 4-class and 3-class task framings to investigate which best supports precancerous detection. Experiments are conducted on a Nigerian colposcopy screening dataset comprising 2,835 images from 282 patients, labelled across four clinically grounded severity grades: Normal, Borderline, Low-grade Precancerous, and High-grade Precancerous. Two deep learning architectures are evaluated across all three task framings under strict patient-level splits. Results show that the hierarchical task framing consistently achieves higher macro-recall across all stages compared to the 4-class and 3-class task framings, with EfficientNet-B3 achieving stable macro-recall of 0.555–0.557 across Stages 1 and 2, versus 0.227 ± 0.020 and 0.366 ± 0.023 for the 4-class and 3-class task framings respectively. These findings suggest that hierarchical task framing offers a more effective and prevention-oriented approach for cervical cancer triage in resource-constrained settings.

Keywords: Cervical Cancer · Colposcopy Image · Hierarchical Classification

1 Introduction

Cervical cancer is the fourth most common cancer in women globally, with over 94% of deaths occurring in low- and middle-income countries (LMICs) where access to trained colposcopists is severely limited [20, 3]. Cervical cancer progresses through a precancerous stage that, if detected early, can be treated before invasive carcinoma develops [1, 14]. This makes precancerous detection the most clinically actionable target for prevention-oriented screening. In LMICs such as

Nigeria, the shortage of specialist clinicians, inadequate vaccination coverage, and disparities in healthcare infrastructure further limit timely diagnosis and management [16, 13].

Deep learning has emerged as a promising tool for automating colposcopy-based cervical lesion detection, with several approaches demonstrating encouraging results [6, 4, 8]. However, most existing methods frame the problem as binary normal-versus-abnormal or cancer-versus-non-cancer classification, or broad multi-class schemes [5, 18]. These task framings collapse or overlook the precancerous class, the category that is most important for prevention, making it difficult to assess whether a model is clinically safe for triage. Whether task framing itself affects the accurate detection of precancerous lesions remains underexplored.

Here, we present a hierarchical task framing for cervical lesion grading that decomposes the classification problem into a cascade of three clinically motivated binary decisions. We evaluate this against conventional 4-class and 3-class task framings using two deep learning architectures on a Nigerian colposcopy screening dataset [15] of 2,835 images from 282 patients, collected at Bowen University Teaching Hospital, Ogbomoso, Nigeria. Results show that the hierarchical task framing consistently achieves higher macro-recall for precancerous lesions across all stages compared to the 4-class and 3-class task framings.

Our main contributions are:

1. We present a hierarchical task framing for colposcopy-based cervical lesion grading that decomposes the problem into three clinically motivated binary decision stages, explicitly preserving the precancerous class.
2. We compare three task framings: hierarchical, 4-class, and 3-class, across two deep learning architectures under strict patient-level splits, demonstrating that hierarchical task framing consistently achieves higher macro-recall than the 4-class and 3-class task framings and contributing evidence for prevention-oriented cervical lesion detection in an LMIC context where such evaluations remain scarce.

2 Related Work

2.1 Deep Learning for Colposcopy-Based Cervical Lesion Detection

Fine-tuned ResNet architectures have reported four-class accuracy of 74.7% under the Lower Anogenital Squamous Terminology (LAST) grading system [5], and acetowhite epithelium segmentation has improved binary High-grade Squamous Intraepithelial Lesion (HSIL) detection by over 10 percentage points [8]. Multimodal approaches combining colposcopy images with HPV status and cytology have reached an AUC of 0.93 for binary lesion classification [21]. Fusing three sequential image types per patient using EfficientNet-B0 and a GRU achieved 91.18% three-class accuracy [4], and combining Swin Transformer feature extraction with a deep learning ensemble further demonstrated the benefit of transformer-based representations for colposcopic classification [6]. Despite these advances, a systematic review and meta-analysis confirmed that high-grade

lesion detection remains the weakest point across published methods [11], motivating the task framing investigation in this work.

2.2 Classification Framework Design and Task Framing

How the classification problem is framed directly shapes what a model learns and which lesions it misses. Binary formulations: normal versus abnormal or cancer versus non-cancer, typically achieve higher aggregate accuracy than multi-class schemes because the decision boundary is simpler, but they collapse the precancerous class that is most actionable for prevention [18, 5]. Broader multi-class schemes preserve more clinical granularity but increase confusion between adjacent grades. It has been demonstrated for gastric cancer staging that different class granularities produce materially different error profiles on the same dataset, establishing that task framing, not architecture alone, determines which lesions are missed [10]. Hierarchical pipelines offer a clinically motivated alternative by decomposing the problem into a cascade of binary decisions that mirror clinical reasoning. Hierarchical medical image classification has been shown to outperform multi-class approaches on biopsy images [9], and similar gains have been reported for ophthalmic disease classification under limited training data, suggesting that hierarchical framing eases class imbalance by focusing model capacity on one decision at a time [2]. However, no study has systematically compared hierarchical, 4-class, and 3-class task framings on colposcopy images with macro-recall as the primary outcome.

2.3 Class Imbalance and High-Grade Lesion Detection

High-grade precancerous lesions are systematically underrepresented in colposcopy datasets, causing models trained on standard loss functions to over-predict the majority class [18, 7]. The clinical cost of this imbalance is asymmetric. A missed HSIL may progress to invasive carcinoma, whereas an unnecessary referral results only in an additional biopsy. Despite this, most published studies report overall accuracy or AUC as primary metrics, obscuring per-class recall and making it difficult to assess model safety for the most critical diagnostic category [11, 17]. By adopting macro-recall as the primary evaluation metric and preserving the precancerous class explicitly across all three task framings, this study addresses a gap that aggregate metrics systematically conceal.

3 Methodology

3.1 Dataset

The Bowen University AI-assisted Cervical Cancer Screening Dataset [15] comprises 3,356 high-resolution colposcopy images from 332 women screened at Bowen University Teaching Hospital, Ogbomoso, Nigeria, captured at four magnification levels (1×, 3×, 5×, 7×) before and after application of 5% acetic

acid, yielding up to eight views per patient. The dataset provides binary labels (Normal/Abnormal), with raw cytology text from per-patient `.txt` files normalised and mapped to a four-class severity hierarchy aligned with the LAST grading system: **Normal** (NILM, reactive), **Borderline** (ASCUS), **Low-grade Precancerous** (LSIL/CIN1), and **High-grade Precancerous** (HSIL/CIN2-3). After data cleaning (removing patients with missing, inadequate or inconsistent labels), 2,835 images from 282 patients were retained. The dataset is imbalanced: Normal (62.1%), Low-grade Precancerous (14.9%), Borderline (13.1%), and High-grade Precancerous (9.9%). Images were resized to 224×224 pixels, normalised using ImageNet statistics, and preprocessed with Wiener filtering, specular reflection removal, and colour normalisation.

3.2 Task Framing

We present a hierarchical task framing that decomposes cervical lesion grading into a cascade of three clinically motivated binary decisions, each trained independently on the relevant patient subset:

- **Stage 1**: Normal vs. Abnormal (all 282 patients)
- **Stage 2**: Borderline vs. Precancerous (abnormal patients only)
- **Stage 3**: Low-grade vs. High-grade Precancerous (precancerous patients only)

This cascade mirrors the sequential reasoning of a colposcopist, first flagging abnormality, then grading severity, then differentiating precancerous grade. By isolating each decision, the hierarchical framing explicitly preserves the precancerous class at every stage. Two conventional task framings serve as baselines: a 4-class framing classifying all four grades simultaneously, and a 3-class framing collapsing the two precancerous grades into a single Precancerous class.

3.3 Model Architectures

Both pipelines share a Swin Transformer (Swin-T) [12] pretrained on ImageNet as a shared feature extraction backbone, producing 768-dimensional representations offline prior to classifier training [6]. The two pipelines differ in their classification heads.

Pipeline A - EfficientNet-B3. Swin-T features are passed to an EfficientNet-B3 [19] backbone with its classifier replaced by a two-layer MLP (Linear $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear(C)). The backbone is initially frozen, with the final two blocks unfrozen after epoch five.

Pipeline B - LSTEDL-CCS Ensemble. Following [6], Swin-T features are processed in parallel by an Autoencoder (AE, 768$\rightarrow$640$\rightarrow$512), a Bidirectional GRU (BiGRU, hidden dim 256), and a Deep Belief Network (DBN, $[512, 256, 128]$). Their outputs (896 dimensions total) are concatenated and passed through a fusion classifier to produce the final prediction.

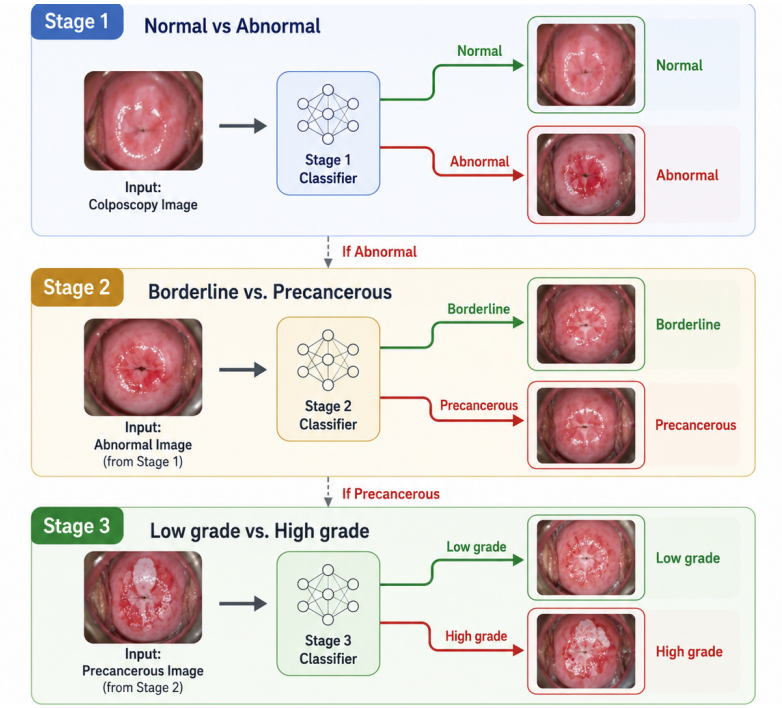


Fig. 1. The three-stage hierarchical task framing.

3.4 Training Strategy

All models train for up to 50 epochs with early stopping on validation macro-recall (patience 10, min delta 10^{-4}), AdamW optimiser ($lr = 10^{-4}$, weight decay 10^{-3}), OneCycleLR scheduling, focal loss ($\gamma = 1.0$, label smoothing 0.1), and class-weighted sampling. Dataset splits are 70/15/15 at patient level using stratified sampling, with WeightedRandomSampler applied to training loaders. The primary metric is macro-averaged recall; secondary metrics are macro-precision, macro-F1, accuracy, and macro-specificity. To assess robustness to initialisation and data split variation, all experiments are repeated across three random seeds (0, 42, 123), with results reported as mean ± standard deviation.

Compute resources. All experiments are conducted on a single NVIDIA A100 GPU (40GB VRAM) on the Alliance Canada Narval HPC cluster. Swin-T feature extraction is performed once per split offline. EfficientNet-B3 models converge in 10–20 epochs; ensemble classifier training on pre-extracted features converges in under 5 minutes. Pipeline B's reliance on pre-extracted features rather than raw image inference makes it substantially faster at inference time than Pipeline A, which processes images end-to-end.

4 Results

4.1 Evaluation Metrics

Given the clinical context of cervical cancer screening, macro-recall (sensitivity) is designated as the primary optimisation target. A missed high-grade lesion carries far greater consequences than falsely flagging a normal case as abnormal, as the latter results only in an unnecessary referral, so model selection and early stopping during training are governed exclusively by validation macro-recall. Formally, macro-recall is computed as the unweighted mean of per-class recall:

$$Macro - Recall = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \tag{1}$$

where K is the number of classes and TP_k and FN_k denote true positives and false negatives for class k. Unweighted averaging gives equal weight to each class regardless of support size, which matters because High-grade Precancerous lesions are substantially outnumbered by Normal cases. Secondary metrics reported include overall accuracy, macro-precision, and macro-F1. For the hierarchical pipeline, each stage is evaluated independently on its patient subset.

4.2 Experiments and Results

Table 1. Test-set performance across task framings and model architectures, reported as mean ± std over 3 seeds (0, 42, 123). Macro-recall is the primary metric.

Pipeline	Task	Classes	Accuracy	Macro-P	Macro-R	Macro-F1
EfficientNet	4-class	4	0.303 ± 0.121	0.227 ± 0.004	0.227 ± 0.020	0.194 ± 0.037
Ensemble	4-class	4	0.206 ± 0.102	0.267 ± 0.045	0.288 ± 0.044	0.171 ± 0.078
EfficientNet	3-class	3	0.472 ± 0.079	0.366 ± 0.032	0.366 ± 0.023	0.359 ± 0.028
Ensemble	3-class	3	0.404 ± 0.106	0.369 ± 0.033	0.381 ± 0.056	0.340 ± 0.063
EfficientNet	Hier. Stage 1	2	0.572 ± 0.028	0.553 ± 0.017	0.555 ± 0.016	0.551 ± 0.019
EfficientNet	Hier. Stage 2	2	0.509 ± 0.073	0.575 ± 0.064	0.557 ± 0.022	0.487 ± 0.069
EfficientNet	Hier. Stage 3	2	0.499 ± 0.127	0.449 ± 0.039	0.474 ± 0.030	0.413 ± 0.070
Ensemble	Hier. Stage 1	2	0.553 ± 0.015	0.519 ± 0.033	0.520 ± 0.033	0.518 ± 0.033
Ensemble	Hier. Stage 2	2	0.485 ± 0.034	0.574 ± 0.041	0.570 ± 0.032	0.482 ± 0.031
Ensemble	Hier. Stage 3	2	0.516 ± 0.156	0.358 ± 0.207	0.518 ± 0.064	0.404 ± 0.159

4-class and 3-class Task Framings. Performance under both 4-class and 3-class framings was modest and highly variable across seeds (Table 1). The ensemble showed substantially higher variance than EfficientNet-B3 under both framings, particularly for accuracy, reflecting instability attributable to class imbalance rather than a consistent architectural advantage for either model.

Hierarchical Task Framing. The hierarchical task framing consistently achieved higher macro-recall than either 4-class and 3-class framing across all stages and both pipelines (Table 1). No single architecture was uniformly superior: EfficientNet-B3 was more stable at Stage 1, while the ensemble attained the highest overall macro-recall at Stage 2, albeit with lower accuracy and greater variance. Stage 3 (Low-grade vs. High-grade Precancerous) was the weakest and least stable point for both models, with the ensemble showing the largest variance observed across all experiments.

End-to-End Hierarchical Pipeline Estimate. To situate the hierarchical results within a single comparable figure, an approximate end-to-end accuracy is derived by chaining mean stage-wise macro-recalls across the test-set class distribution (285 Normal, 50 Borderline, 73 Low-grade, 42 High-grade; $n = 450$):

$$\hat{A} = r_1 \cdot p_N + r_1 r_2 (p_B + p_P) + r_1 r_2 r_3 (p_L + p_H)$$

where r_1, r_2, r_3 are mean stage macro-recalls and p_N, p_B, p_L, p_H are test-set class proportions. This formulation assumes independence between stages and should be read as an approximation rather than an exact figure. EfficientNet achieves an estimated end-to-end accuracy of approximately 0.50, compared with 0.48 for the ensemble, both exceeding the mean 4-class accuracy for their respective model families (0.30 and 0.21).

5 Discussion

The results suggest that task framing meaningfully influences precancerous detection. Decomposing the four-class grading problem into a cascade of binary decisions consistently raised macro-recall above either 4-class or 3-class task framing on the same limited dataset. This improvement is largely structural. Each stage resolves one well-defined distinction at a time, reducing inter-class confusion. Whether this pattern holds on larger, more diverse datasets remains an open question.

Stage 3 (Low-grade vs. High-grade Precancerous) warrants careful interpretation. Both models showed marginal discrimination (macro-recall 0.474 ± 0.030 for EfficientNet-B3, 0.518 ± 0.064 for the ensemble), but the failure mode differs. EfficientNet-B3 is consistently limited, while the ensemble is highly unstable, with the largest variance observed across all experiments (macro-precision std 0.207). Both are consistent with a single cause, the boundary between low-grade and high-grade dysplasia is subtle and conventionally confirmed histologically rather than visually, suggesting genuine visual ambiguity rather than a correctable modelling failure. This is a notable tension: the hierarchical framing improves detection of precancerous status overall, yet offers the least improvement precisely at the grading decision (Low-grade vs. High-grade) most consequential for treatment planning, indicating that task decomposition alone cannot overcome the underlying visual ambiguity at this boundary.

None of the evaluated framings are close to clinical deployment thresholds, and no single architecture is uniformly superior across stages. A macro-recall of 0.570 at the best-performing hierarchical stage, attained by the ensemble at Stage 2, means roughly four in ten precancerous cases would still be missed or misclassified under the most favourable configuration tested, and this came with lower accuracy and greater variance than EfficientNet-B3's comparable Stage 2 result (0.557 ± 0.022). This variance makes the ensemble a less reliable deployment choice despite occasionally matching or exceeding EfficientNet-B3 on the primary metric. The contribution of this work is therefore not a deployable triage tool, but evidence that task framing itself is a meaningful lever for improving precancerous detection, one that should be combined with larger, multi-site, histopathologically confirmed datasets before clinical translation can be considered, and underscores the importance of reporting variance for models intended for clinical triage.

Limitations. The dataset is drawn from a single institution using a single colposcopy device, limiting generalisability across clinical sites and equipment types. Labels derive from colposcopic images and cytology reports rather than histopathological confirmation, introducing noise at the Low-grade versus High-grade boundary where both models struggle most. Class definitions also vary across studies, making direct numerical comparison with published results unreliable.

6 Impact in Resource-Constrained Settings

The proposed pipeline is designed with LMIC deployment in mind. Swin-T feature extraction is performed offline as a one-time preprocessing step, decoupling computational cost from real-time inference; Pipeline B's reliance on pre-extracted features rather than raw image processing makes it substantially faster at inference than Pipeline A, which processes images end-to-end. The hierarchical framing further supports resource-constrained deployment through early termination, as a Stage 1 Normal prediction requires no further computation, reserving resources for abnormal cases requiring grading.

7 Conclusion

This study presents a hierarchical task framing for cervical lesion grading and compares it against 4-class and 3-class task framings on a Nigerian colposcopy dataset under strict patient-level splits, evaluated across three random seeds. The hierarchical task framing consistently achieves higher macro-recall across all stages, though no single architecture is uniformly superior. EfficientNet-B3 is more stable overall, while the ensemble attains the highest mean macro-recall at Stage 2 with greater variance. Low-grade vs. High-grade Precancerous discrimination at Stage 3 remains the critical bottleneck, with macro-recall of 0.474 ± 0.030 for EfficientNet-B3 and 0.518 ± 0.064 for the ensemble, reflecting both visual ambiguity in the colposcopic modality and potential label noise at

this boundary. Future work should target this boundary through histologically confirmed labels, multimodal fusion incorporating HPV typing, and validation across multiple clinical sites and device types.

References

1. Allogmani, A., Mohamed, R., Al-Shibly, N., Ragab, M.: Enhanced cervical precancerous lesions detection and classification using Archimedes optimization algorithm with transfer learning. Scientific Reports **14** (2024). https://doi.org/10.1038/s41598-024-62773-x
2. An, G., Akiba, M., Omodaka, K., Nakazawa, T., Yokota, H.: Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. Scientific Reports **11**(1) (2021). https://doi.org/10.1038/s41598-021-83503-7
3. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **74**(3), 229–263 (2024). https://doi.org/10.3322/caac.21834
4. Chen, X., Pu, X., Chen, Z., Li, L., Zhao, K., Liu, H., Zhu, H.: Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. Cancer Medicine **12**, 8690–8699 (2023). https://doi.org/10.1002/cam4.5581
5. Cho, B., Choi, Y., Lee, M., Kim, J., Son, G., Park, S., et al.: Classification of cervical neoplasms on colposcopic photography using deep learning. Scientific Reports **10**, 13652 (2020). https://doi.org/10.1038/s41598-020-70490-4
6. Himabindu, D., Lydia, E., Rajesh, M., Ahmed, M., Ishak, M.: Leveraging Swin transformer with ensemble of deep learning models for cervical cancer screening using colposcopy images. Scientific Reports **15** (2025). https://doi.org/10.1038/s41598-025-90415-3
7. Kaur, H., Sharma, R., Kaur, J.: Comparison of deep transfer learning models for classification of cervical cancer from pap smear images. Scientific Reports **15** (2025). https://doi.org/10.1038/s41598-024-74531-0
8. Kim, J., Park, C., Kim, S., Cho, A.: Convolutional neural network-based classification of cervical intraepithelial neoplasias using colposcopic image segmentation for acetowhite epithelium. Scientific Reports **12**, 17228 (2022). https://doi.org/10.1038/s41598-022-21692-5
9. Kowsari, K., Sali, R., Ehsan, L., Adorno, W., Ali, A., Moore, S., Amadi, B., Kelly, P., Syed, S., Brown, D.: HMIC: Hierarchical medical image classification, a deep learning approach. Information **11**(6), 318 (2020). https://doi.org/10.3390/info11060318
10. Lima, G., Coimbra, M., Dinis-Ribeiro, M., Libanio, D., Renna, F.: Analysis of classification tradeoff in deep learning for gastric cancer detection pp. 2177–2180 (2022). https://doi.org/10.1109/embc48229.2022.9871040
11. Liu, L., Liu, J., Su, Q., Chu, Y., Xia, H., Xu, R.: Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: a systematic review and meta-analysis. EClinicalMedicine **80**, 102992 (2024). https://doi.org/10.1016/j.eclinm.2024.102992
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
13. Mafiana, J.J., Dhital, S., Halabia, M., Wang, X.: Barriers to uptake of cervical cancer screening among women in Nigeria: a systematic review. African Health Sciences **22**(2), 295–309 (2022). https://doi.org/10.4314/ahs.v22i2.33
14. Mathivanan, S., Francis, D., Srinivasan, S., Khatavkar, V., P, K., Shah, M.: Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. Scientific Reports **14** (2024). https://doi.org/10.1038/s41598-024-61063-w
15. Ogunlaja, A.O., Adeleke, O.T., Ughonu, P.C., Oyelami, M.O., Ano-Edward, G., Bakare, Y.T., Bobo, I.T., Atanda, G.O., Oyebade, I.A., Afolabi, A.O., Adeniji, O.D., Bello, G.O., Okunoye, O.O., Ladoye, O.O.: A multi-magnification digital colposcopy image datasets for cervical cancer screening in Nigeria. Data in Brief **65**, 112633 (Apr 2026). https://doi.org/10.1016/j.dib.2026.112633, pMID: 41852843; PMCID: PMC12992497
16. Ola, O., Okunowo, A., Habeebu, M., Miao Jonasson, J.: Clinical and non-clinical determinants of cervical cancer mortality: A retrospective cohort study in Lagos, Nigeria. Frontiers in Oncology **13**, 1105649 (2023). https://doi.org/10.3389/fonc.2023.1105649
17. Qin, D., Bai, A., Xue, P., Seery, S., Wang, J., Mendez, M., Li, Q., Jiang, Y., Qiao, Y.: Colposcopic accuracy in diagnosing squamous intraepithelial lesions: a systematic review and meta-analysis of the IFCPC 2011 terminology. BMC Cancer **23**(1), 187 (2023). https://doi.org/10.1186/s12885-023-10648-1
18. Sarhangi, H., Beigifard, D., Farmani, E., Bolhasani, H.: Deep learning techniques for cervical cancer diagnosis based on pathology and colposcopy images. arXiv **abs/2310.16662** (2023). https://doi.org/10.48550/arxiv.2310.16662
19. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). pp. 6105–6114 (2019)
20. World Health Organization: Global cancer statistics 2022. Tech. rep., World Health Organization (2022), https://www.who.int/publications/i/item/9789240074590
21. Yuan, C., Yao, Y., Cheng, B., Cheng, Y., Li, Y., Li, Y., Liu, X., Cheng, X., Xie, X., Wu, J., Wang, X., Lu, W.: The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. Scientific Reports **10**, 11639 (2020). https://doi.org/10.1038/s41598-020-68252-3