Conformal Prediction and Monte Carlo Inference for Addressing Uncertainty in Cervical Cancer Screening

Christopher Clark[1], Scott Kinder[1], Didem Egemen[2], Brian Befano[3], Kanan Desai[2], Syed Rakin Ahmed[4], Praveer Singh[1], Ana Cecilia Rodriguez[2], Jose Jeronimo[2], Silvia De Sanjose[5], Nicolas Wentzensen[2], Mark Schiffman[2], and Jayashree Kalpathy-Cramer[1]

[1] University of Colorado School of Medicine, USA
{christopher.w.clark, scott.kinder, praveer.singh, jayashree.kalpathy-cramer}@cuanschutz.edu
[2] National Cancer Institute, USA
{didem.egemen, kanan.desai, rodriguezac2, jose.jeronimo, wentzenn, schiffmm}@nih.gov
[3] Information Management Services, USA
befanob@imsweb.com
[4] Harvard University, USA
syedrakin_ahmed@fas.harvard.edu
[5] Cancer Epidemiology Research Programme, Spain
s.sanjose@iconcologia.net

Abstract. In the medical domain, where a misdiagnosis can have life-altering ramifications, understanding the certainty of model predictions is an important part of the model development process. However, deep learning approaches suffer from a lack of a native uncertainty metric found in other statistical learning methods. One common technique for uncertainty estimation is the use of Monte-Carlo (MC) dropout at training and inference. Another approach is Conformal Prediction for Uncertainty Quantification (CUQ). This paper will explore these two methods as applied to a cervical cancer screening algorithm currently under development for use in low-resource settings. We find that overall, CUQ and MC inference produce similar uncertainty patterns, that CUQ can aid in model development through class delineation, and that CUQ uncertainty is higher when the model is incorrect, providing further fine-grained information for clinical decisions. Code available here

Keywords: Conformal Prediction · Computer Vision · Cervical Cancer

1 Introduction

Cervical cancer is the fourth-leading cancer in women worldwide, and poses a significant threat in lower and middle-income countries due primarily to inequalities in access to vaccination, screening, and treatment services [16]. A common

screening method, colposcopies, require expertise to properly administer and interpret results, and this can be a challenge in some areas of the world [10]. The *hPv-Automated Visual Examination* (PAVE) project has developed an *Automated Visual Evaluation* (AVE) algorithm, a *deep learning* (DL) model for cervical cancer screening [5]. This algorithm was designed to act in conjunction with HPV genotyping to triage the risk of HPV-positive individuals, perhaps with additional methods, such as *Visual Inspection with Acedic Acid* (VIA) [5,1]. The current, three-class model showed improved performance relative to the two-class version, as many of the "Normal" images were reclassified as "Gray Zone", providing an intermediary between "Normal" and "Precancer+" [5]. However, this middle class suffers from substantial interrater variability in diagnosis [13]. It is model uncertainty and its relationship to model performance, particularly around the "Gray Zone" class, that we explore in this work through the use of a metric derived from *Monte Carlo* (MC) inference and another technique, *Conformal Prediction for Uncertainty Quantification* (CUQ).

In this work, we first determine the relationship between model uncertainty and performance using different CUQ algorithms. Next, we compare CUQ to the results of the uncertainty as determined through MC inference. Finally, we hope to understand the aleatoric uncertainty surrounding the "Gray Zone" class. Though other research groups have approached the uncertainty problem in DL with CUQ in the medical domain, such as in skin lesion classification [12] and prostate cancer [15], not all results have been positive, as in [14]. Our contribution will be the exploration of CUQ for determining the effect of ground-truth categorization on model uncertainty and better understanding misclassifications, especially "Normal" to "Precancer+" or vice versa, in the cervical cancer domain for applications in low-resource settings.

2 Methods

2.1 Conformal Prediction Overview

For classification, our goal is to develop a model, $\hat{f}_y(x)$, which estimates the quantity $\mathbb{P}[Y = y|X = x]$, where y is the label and x is the datum, with outputs in Δ^K, the K-simplex.

To understand the model uncertainty, we will construct a conformal prediction set for a test point x_{test}, $\hat{C}(x_{test}) \subseteq \mathcal{Y}$, where $\mathcal{Y}$ is all our possible classes (i.e., $|\mathcal{Y}| = K$), such that $\mathbb{P}[y_{test} \in \hat{C}(x_{test})] \geq 1-\alpha$. We call $1-\alpha$ the (empirical) *coverage* and α is the *error rate* [2].

2.2 Least Ambiguous Set-Valued Classifier

We will briefly describe two conformal prediction algorithms, beginning with *Least Ambiguous Set-Valued Classifier* (LAC) [2].

We will divide our data $\mathcal{X}$ into three sets, $\mathcal{X}_{train}$, $\mathcal{X}_{calibration}$, and $\mathcal{X}_{test}$, where $\mathcal{X}_{train}$ is our standard training set used to train the model $\hat{f}$, $\mathcal{X}_{calibration}$

is a calibration set to prepare for our conformal predictions, and $\mathcal{X}_{test}$ is the set of data we wish to construct conformal predictions for. Let n_{cal} be the number of calibration points.

Now, we introduce a *score* function, $s(x, y)$ which tells us how well the model is performing. The LAC algorithm uses the probability of that specific class. As in, if $\hat{f}(x)_y = [p_0, \ldots, p_{K-1}]$, we can take:

$$s(x, y) = 1 - \hat{f}(x)_{y_i}, y_i = \text{Index of correct class}$$

For each element of our calibration set $\mathcal{X}_{cal}$, we repeat the above process, giving us $\{s_1, \ldots, s_{n_{cal}}\}$, from which we calculate the *quantile*:

$$\hat{q} = quantile\left(\{s_1, \ldots, s_{n_{cal}}\}; \frac{\lceil (1-\alpha)(n_{cal}+1) \rceil}{n_{cal}}\right)$$

From this, we can construct our $\hat{C}(x_{test})$ as:

$$\hat{C}(x_{test}) = \{y : s(x_{test}, y_{test}) \leq \hat{q}\} = \{y : \hat{f}(x_{test})_y \geq 1 - \hat{q}\}$$

Since we don't have y_{true} for our test point, we are choosing all the indices of $\hat{f}(x_{test})$ with a value greater than $1 - \hat{q}$.

2.3 Adaptive Prediction Sets

For our second algorithm, the *Adapative Prediction Set* (APS) version of conformal prediction, we begin by changing our score function [2]. Now, we will take all the softmaxed output scores and arrange them by size, taking us from $\hat{f}(x)$ to $\pi(x)$. The correct class will appear at some index k of the rearranged probability vector, $y = \pi_k(x)$, and we sum up to this index along $\pi(x)$:

$$s(x, y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}$$

We create our $\hat{q}$ same as above, and then our prediction set is created by:

$$\hat{C}(x_{\text{test}}) = \{\pi_1(x_{test}), \ldots, \pi_k(x_{test})\}, \quad k = \sup\left\{k' : \sum_{j=1}^{k'} \hat{f}(x)_{\pi_j(x)} < \hat{q}\right\}$$

These prediction sets carry a native measure of model uncertainty, the number of classes included in each set, denoted the *length*, which we will use as our uncertainty metric [2].

2.4 Uncertainty Estimation with Monte Carlo Dropout and Inference

Dropout is used as a regularization tool to improve generalizability by preventing the model from "memorizing" the data and addressing the problem of epistemic uncertainty [20]. As an additional outcome, by leaving dropout on during inference and running several inferences, the result mimics an *ensemble* of models [7] and allows for uncertainty quantification [6]. Thus, in addition to normal inference (i.e., with all stochastics turned off), we run 50 inferences per datum with dropout left on to generate our MC predictions. From these 50 predictions, we find the *expected value* of each prediction, $\mathbb{E}[\mathbf{p}] = \sum_{i=0}^{2} ip_i$. We define our uncertainty with the MC method as the *coefficient of variation* (CoV), $\frac{\sigma}{\mu}$, of these expected values.

3 Model Development and Data Description

The use of AI for cervical cancer screening has been explored before, as in [21] with dual-stain cytology. However, collection, transportation, and analysis of cytological samples requires significant infrastructure and expertise [10]. Alternative screening methods, such as VIA have been proposed and explored, but suffer from high subjectivity and variability [4,19]. DL has also been applied to cervical images themselves as a screening methodology, but a lack of performance on, or absence of, a held-out test set and the inability to maintain performance in different settings demands additional development [8,17,22,18]. With these considerations, the current, best performing AVE model is a three-class model trained to classify cervical images taken during a colposcopy into "Normal", "Gray Zone", and "Precancer+", denoting a normal cervix, unsure/not sufficiently advanced to determine, and already or likely to result in cancer [9]. To improve repeatability, *Monte-Carlo* (MC) dropout was introduced to positive effects [1,11]. With the inclusion of a dropout layer, a notion of uncertainty in model predictions can be measured [3] as described in the **Methods** section. We will be using this model, as well as the closest-performing two-class model, for our investigation.

The final, labeled dataset has 9,462 women from five studies conducted in Costa Rica, the US, and the Netherlands, for a total of 17,013 images. Each study has its own particulars which can be found in the supplementary material for [5], but we highlight that the images are of cervices captured by a standard cervicscope or a Nikon digital single-lens reflex (DSLR) camera during a colposcopy and resized to 224×224 for the model. These were divided into training, validation, test 1 and test 2, splitting on patient level, resulting in a *data percentage split* of $\approx 33/6/51/10$. The AVE study is using test 2 as the out-of-distribution dataset, and so we maintain this here. We are using a calibration/testing division of 20/80 of test 2, as this was used as the out-of-distribution set after final model selection. This has 1348 images. Regarding the difference in ground-truth determination in the two models, all the "Gray Zone" images in the three-class model were originally given a "Normal" ground truth in the two-class model [5].

4 Experiment

First, an appropriate α value and algorithm choice needs to be made. We have run both LAC and APS with $\alpha = 0.05, 0.1$, and 0.2 and decided that LAC with $\alpha = 0.1$ resulted in the most reasonable set sizes. This determination was made by balancing variation in the prediction set sizes and empirical coverage. Some algorithm and α combinations created length-0 sets, indicating a failure, or gave mostly length-3 sets, denying any nuanced analysis. So, we will display here the results of LAC and APS with $\alpha = 0.1$, but we will include more in the Supplementary Material.

Task 1: Accuracy and Conformal Prediction Set Length We subset our conformal results based on *correct*, *incorrect*, *single-class misclassifications* (SC), i.e., "Normal" to "Gray Zone", and *two-class misclassifications* (TC), i.e., "Normal" to "Precancer+". Though SC errors are an issue, given the risks with missing a cancerous lesion or performing an unnecessary biopsy, it is these TC errors that are most concerning. We also look at each ground truth individually, taking the average prediction set length of the images correctly classified as that class and comparing it to the images incorrectly classified as that class.

Task 2: Relationship between MC Uncertainty and CUQ We calculate the Spearman's correlation coefficient [23] between the average conformal prediction set length and coefficient of variation of the expected values of the MC inferences and provide a distribution graphic of the coefficients of variation color-coded by conformal prediction length.

Task 3: Role of "Gray Zone" in Three and Two-Class Models We find the average conformal prediction set length of the images with each ground truth from the three-class model and run a t-test of these averages to determine if the model is less certain of particular ground truths, as we expect it is uncertain of "Gray Zone" images. Then, we pivot to the two-class model to determine the uncertainty around the images that in the three-class model were given a ground truth of "Gray Zone" and perform a series of t-tests on these averages to validate through uncertainty the decision to use a three-class model over a two-class version.

5 Results

5.1 Task 1: Accuracy and Conformal Prediction Set Length

From Figure 1, we compare the diagonals, (i.e., predicted "Normal" and ground-truth "Normal", etc.) and we see they are always less than or equal to the rest of values in the same predicted column. In Table 1, we also see that though the correct predictions have a lower average conformal prediction set length, the single-class misclassifications have larger average conformal prediction set lengths than the two-class misclassifications, which is surprising. However, looking at the confusion matrices, we see strong uncertainty for both algorithms when the model predicts incorrectly an image as belonging to the "Gray Zone" class, showing how the model struggles with this class and explaining why the

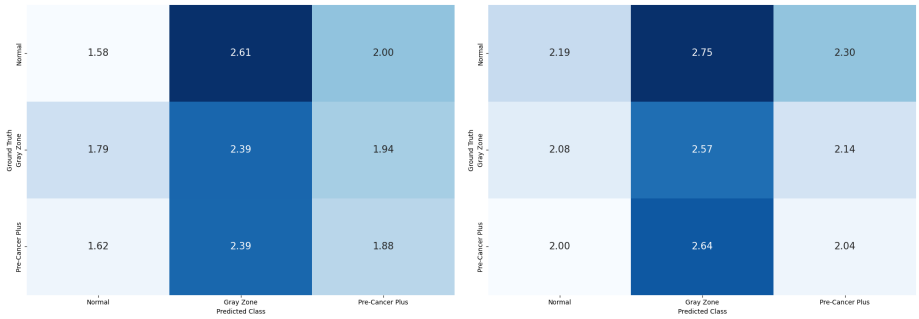

Fig. 1: Confusion Matrix of Average Conformal Prediction Length for LAC (Left) and APS (Right) with α=0.1

Table 1: T-Test for Average Prediction Length, μ_{PL}, Comparison for Correct, Incorrect, Single, Two-Class Misclassification and Per Ground Truth Correct or Not (95% Confidence). * indicates $p < 0.05$.

$\mu_{\text{PL},1}$ vs $\mu_{\text{PL},2}$	LAC	APS	p_{LAC}	p_{APS}
Corr vs Incor	1.78 ± 0.05 vs 2.38 ± 0.06	2.26 ± 0.04 vs 2.57 ± 0.05	*	*
Corr vs SC	1.78 ± 0.05 vs 2.43 ± 0.06	2.26 ± 0.04 vs 2.60 ± 0.05	*	*
Corr vs TC	1.78 ± 0.05 vs 1.93 ± 0.13	2.26 ± 0.04 vs 2.24 ± 0.14	0.16	0.81
SC vs TC	2.43 ± 0.06 vs 1.93 ± 0.13	2.60 ± 0.05 vs 2.24 ± 0.14	*	*
GT 0: Corr vs Incor	1.58 ± 0.05 vs 2.54 ± 0.07	2.19 ± 0.04 vs 2.70 ± 0.05	*	*
GT 1: Corr vs Incor	2.39 ± 0.07 vs 1.90 ± 0.10	2.57 ± 0.07 vs 2.12 ± 0.08	*	*
GT 2: Corr vs Incor	1.88 ± 0.13 vs 2.31 ± 0.12	2.04 ± 0.10 vs 2.57 ± 0.11	*	*

single-class misclassifications have larger prediction set sizes. Figure 1 in the Supplementary Material displays the confusion matrices for LAC with $\alpha = 0.2$ and 0.05. These analyses demonstrate that generally, when the model is wrong, it is also uncertain, one of the key findings that if it were not true, would frustrate attempts to use uncertainty clinically.

5.2 Task 2: Relationship between MC Uncertainty and CUQ

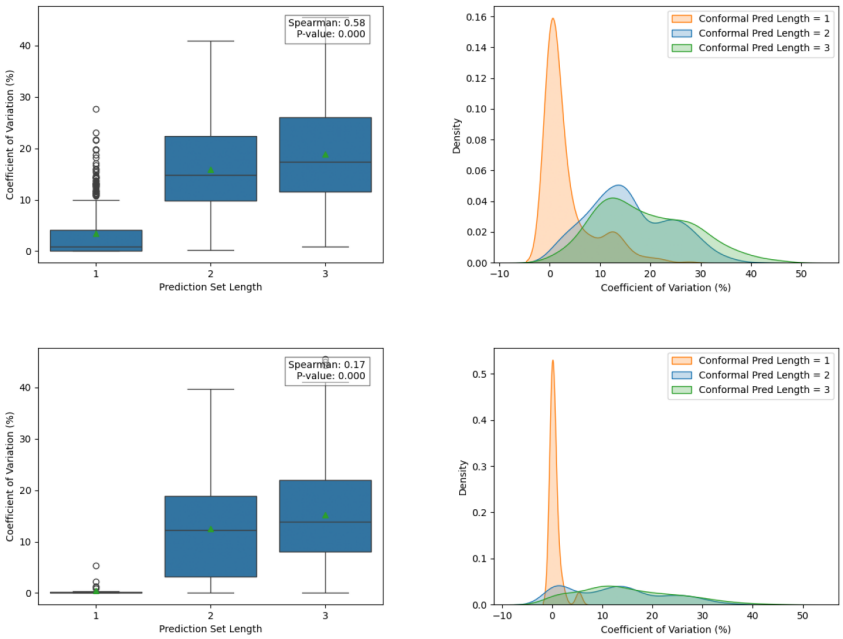

Fig. 2: Box-and-Whisker Plot of Conformal Prediction Length vs Coefficient of Variation (Left) and Distribution of Coefficient of Variation Color-Coded by Conformal Prediction Set Length (Right) for LAC (Top) and APS (Bottom) with $\alpha = 0.1$

From Figure 2, we demonstrate that as the conformal prediction set length increases, so do the coefficients of variation, and that there is a strong correlation between the two in the LAC case. However, though this relationship holds for APS, it is not quite as strong. From these two points, we establish a connection between MC uncertainty and these CUQ algorithms. Figure 2 in the Supplementary Material shows the same figure for LAC with $\alpha = 0.2$ and 0.05.

5.3 Task 3: Average Conformal Conformal Prediction Set Lengths by Ground Truth

From Table 2, we see that the "Gray Zone" and "Precancer+" classes have higher average prediction set lengths than "Normal" for the LAC. However, between the "Gray Zone" and "Precancer+", the difference isn't statistically significant at the $p = 0.05$ level. With the APS algorithm, we find that the "Normal" and "Precancer+" averages are closer and the differences are not statistically significant.

Table 2: T-Test for Average Prediction Length, μ_{PL}, Comparison by Ground-Truth Class (95% Confidence). * indicates $p < 0.05$.

$\mu_{\text{PL},1}$ vs $\mu_{\text{PL},2}$	LAC	APS	p LAC	p APS
Normal vs GZ	1.89 ± 0.05 vs 2.24 ± 0.07	2.35 ± 0.04 vs 2.43 ± 0.06	*	*
Normal vs PC+	1.89 ± 0.05 vs 2.13 ± 0.11	2.35 ± 0.04 vs 2.34 ± 0.09	*	0.82
GZ vs PC+	2.24 ± 0.07 vs 2.13 ± 0.11	2.43 ± 0.06 vs 2.34 ± 0.09	0.07	0.10

5.4 Task 3: "Gray Zone" in the Two-Class Model

Table 3 shows the results of statistical tests on the average conformal prediction set length of the images in the two-class model that were re-classified as "Gray Zone" and compares them to the overall average prediction set length for the remaining images given "Normal" and "Precancer+" in the three-class model.

Table 3: T-Test for Average Conformal Prediction Lengths Compared to the "Gray Zone", μ_{GZ}, by LAC with $\alpha = 0.1$ in the Two-Class Model (95% Confidence). * indicates $p < 0.05$.

μ_{GZ} vs μ_{PL}	LAC	p
GZ	1.51 ± 0.06	(ref.)
Overall Inc GZ	1.31 ± 0.03	*
Overall Exc GZ	1.26 ± 0.03	*
Normal	1.23 ± 0.03	*
PC+	1.46 ± 0.09	0.28

With the LAC algorithm, we see a clear and statistically significant difference in the average conformal prediction set lengths of the "Normal" and the "Gray Zone" images, showing that the model is uncertain about this subset of images. This pattern also holds for the APS algorithm and this table with the APS values can be found in the Supplementary Material Table 1 for space considerations.

6 Discussion

In this paper, we have explored conformal prediction as a means of measuring model uncertainty in a specific case of cervical cancer screening using the AVE model. Focusing on the LAC algorithm, we were able to see significant differences in the uncertainty between correct predictions and misclassifications. With the APS version, we still saw this, but it was not as strong. We also see a connection between MC inference uncertainty and CUQ through the Spearman correlation coefficients in Figure 2, but the relationship is not as strong with APS. From [1], we see that experiments showed better performance with three classes, having taken many "Normal" images and reclassifying them as "Gray Zone" and having the model predict this class, as well, and with the LAC algorithm, the model is less certain of these images. The difference in the three analyses with the LAC and APS algorithms can be explained by the overall larger prediction set sizes by the APS algorithm, which do not allow for as nuanced of an analysis.

Though these two techniques, CUQ and MC inference, deliver comparable results, their implementations require careful thought about the kind of resources available for the user. MC runs require the image to be passed through the model several times, adding to inference time. CUQ bypasses this, as the user only needs to store the $\hat{q}$ on the device and the rest of the operations do not require significant computational resources. Further, using CUQ to determine partitions of data into classes is helpful when there is not an *a priori*, or obvious, way to do so, aiding in model development. Additionally, the uncertain images could then be removed to see if the model improves and/or further analyzed to determine *why* the model is uncertain of them beyond the ground truth, perhaps exposing a flaw in their capture, like blur, or the presence of obfuscating mucus or blood, allowing the clinician to retake the image under different circumstances. However, the drawback to this method is that the choice of alpha, type of CUQ algorithm, creation of the calibration set, etc., can have a marked effect on the outcomes.

References

1. Syed Rakin Ahmed, Brian Befano, Andreanne Lemay, Didem Egemen, Ana Cecilia Rodriguez, Sandeep Angara, Kanan Desai, Jose Jeronimo, Sameer Antani, Nicole Campos, et al. Reproducible and clinically translatable deep neural networks for cervical screening. *Scientific reports*, 13(1):21772, 2023.
2. Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
3. Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, Eline Kooi, Aad Van Der Lugt, and Marleen De Bruijne. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pages 32–41. Springer, 2020.

4. R Catarino, S Schäfer, P Vassilakos, P Petignat, and M Arbyn. Accuracy of combinations of visual inspection using acetic acid or lugol iodine to detect cervical precancer: a meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(5):545–553, 2018.
5. Kanan T Desai, Brian Befano, Zhiyun Xue, Helen Kelly, Nicole G Campos, Didem Egemen, Julia C Gage, Ana-Cecilia Rodriguez, Vikrant Sahasrabuddhe, David Levitz, et al. The development of "automated visual evaluation" for cervical cancer screening: the promise and challenges in adapting deep-learning for clinical testing. *International journal of cancer*, 150(5):741–752, 2022.
6. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
7. Kazuyuki Hara, Daisuke Saitoh, and Hayaru Shouno. Analysis of dropout learning regarded as ensemble learning. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 72–79. Springer, 2016.
8. Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute*, 111(9):923–932, 2019.
9. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
10. HC Kitchener, PE Castle, and JT Cox. Chapter 7: Achievements and limitations of cervical cytology screening. vaccine [internet]. 2006 [acceso 23/09/2019]; 24 (suppl 3): S3/63-70.
11. Andreanne Lemay, Katharina Hoebel, Christopher P Bridge, Brian Befano, Silvia De Sanjosé, Didem Egemen, Ana Cecilia Rodriguez, Mark Schiffman, John Peter Campbell, and Jayashree Kalpathy-Cramer. Improving the repeatability of deep learning models with monte carlo dropout. *npj Digital Medicine*, 5(1):174, 2022.
12. Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016, 2022.
13. Kathrine Dyhr Lycke, Jayashree Kalpathy-Cramer, Jose Jeronimo, Silvia De Sanjose, Didem Egemen, Marta Del Pino, Jenna Marcus, Mark Schiffman, and Anne Hammer. Agreement on lesion presence and location at colposcopy. *Journal of lower genital tract disease*, 28(1):37–42, 2024.
14. Hendrik Mehrtens, Tabea Bucher, and Titus J Brinker. Pitfalls of conformal predictions for medical image classification. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 198–207. Springer, 2023.
15. Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, Marco Capuccini, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Emiel AM Janssen, Anders Blilie, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications*, 13(1):7761, 2022.
16. World Health Organization. Cervical cancer. `https://www.who.int/news-room/fact-sheets/detail/cervical-cancer`, 2024. Accessed: 2024-06-19.

17. Anabik Pal, Zhiyun Xue, Brian Befano, Ana Cecilia Rodriguez, L Rodney Long, Mark Schiffman, and Sameer Antani. Deep metric learning for cervical image classification. *IEEE Access*, 9:53266–53275, 2021.
18. Saritha Shamsunder, Archana Mishra, Anita Kumar, Rajni Beriwal, Charanjeet Ahluwalia, and Sujata Das. Diagnostic accuracy of artificial intelligence algorithm incorporated into mobileodt enhanced visual assessment for triaging screen positive women after cervical cancer screening. 2022.
19. Shannon L Silkensen, Mark Schiffman, Vikrant Sahasrabuddhe, and John S Flanigan. Is it time to move beyond visual inspection with acetic acid for cervical cancer screening?, 2018.
20. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
21. Nicolas Wentzensen, Bernd Lahrmann, Megan A Clarke, Walter Kinney, Diane Tokugawa, Nancy Poitras, Alex Locke, Liam Bartels, Alexandra Krauthoff, Joan Walker, et al. Accuracy and efficiency of deep-learning–based automation of dual stain cytology in cervical cancer screening. *JNCI: Journal of the National Cancer Institute*, 113(1):72–79, 2021.
22. Zhiyun Xue, Akiva P Novetsky, Mark H Einstein, Jenna Z Marcus, Brian Befano, Peng Guo, Maria Demarco, Nicolas Wentzensen, Leonard Rodney Long, Mark Schiffman, et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. *International Journal of Cancer*, 147(9):2416–2423, 2020.
23. Jerrold H Zar. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7, 2005.