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

No Author Given

No Institute Given

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.

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 model was 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. This algorithm was designed to act in conjunction with HPV genotyping to triage the risk of HPV-positive individuals, perhaps in conjunction with other methods, such as *Visual Inspection with Acedic Acid*, or VIA [5], [1]. An issue identified during the development of this model was a lack of repeatability of the model output, and this was deemed a critical component of subsequent model selection processes [1], [11]. In an effort to increase 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]. Further, though the three-class model showed improved performance relative to the two-class version, the "Gray Zone" class as an intermediary between "Normal" and "Precancer+" suffers from substantial interrater variability in diagnosis [13]. It is model uncertainty and its relationship to model performance and the "Gray Zone" class that we explore in this work through the use of MC dropout and another technique, *Conformal Prediction for Uncertainty Quantification* (CUQ).

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]. In a model that is quite "certain" (to be defined later), output classes (or the probability vectors) should remain roughly the same over the inferences, and the differences can be analysed for variation. CUQ provides a different approach, outputting a *prediction set* which includes all the classes required to achieve a pre-determined level of *coverage*. These prediction sets carry a native measure of model uncertainty, the number of classes included in each set, denoted the *length* [2].

In this work, we first attempt to determine the relationship between model uncertainty and accuracy 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

In this section, we will focus briefly on the development of the model, describe the two CUQ algorithms we will use, and describe the experimental setup.

2.1 Model Development

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]. The AVE model was developed to answer these issues and the best model was found to be a three-class *DensNet121* [9] model with images resized to 224 × 224. We will be using this model, as well as the closest-performing two-class model, for our investigation [5].

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 cerviscope or a Nikon digital single-lens reflex (DSLR) camera during a colposcopy. 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. Regarding the difference in ground-truth determination, all the "Gray Zone" images in the three-class model were originally given a "Normal" ground truth in the two-class model [5].

2.2 Conformal Prediction Overview

We begin with a brief review of conformal prediction as applied to classification [2].

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

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

2.3 Least Ambiguous Set-Valued Classifier

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

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 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:

$$s(x, y) = 1 - \hat{f}(x)_{y_i}$$

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 calulate 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 scores greater than $1 - \hat{q}$.

2.4 Adaptive Prediction Sets

For our second algorithm, the *Adapative Prediction Set* (APS) version of conformal prediction, we begin by changing our score function. 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:

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

2.5 Experimental Setup

In addition to normal inference (i.e., with all stochastics turned off), we ran 50 inferences per datum with dropout to generate our MC predictions. From these 50 predictions, we find the *expected value* of the prediction, $\sum_{i=0}^{2} ip_i$. This allows us to get a real number from each MC prediction. We define our uncertainty with the MC method as the *coefficient of variation*, σ/μ, of these expected values. For each datum we have two measures of uncertainty, the conformal prediction set length and the coefficient of variation of the 50 expected values of the MC predictions.

First, an appropriate α value and algorithm choice needs to be made to display our results. 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. So, we will display here the results of LAC and APS with $\alpha = 0.1$, but we will include more about LAC with $\alpha = 0.05$ and 0.2 in the Supplementary Material.

Finally, 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. In summary, we are comparing two techniques of uncertainty quantification, MC and CUQ, and two variations of CUQ, LAC and APS, for a total of three methods.

3 Results

We will organize our results around our three tasks: exploring the model's confidence surrounding misclassifications, comparing CUQ and MC inference, and determining the role of the "Gray Zone" class in our two and three-class models.

3.1 Relationship between conformal prediction set lengths and accuracy

We subset our conformal results based on *correct*, *incorrect*, *single-class misclassifications* (SC), and *two-class misclassifications* (TC). We then compare the average prediction set lengths in each instance in Figure 1 and Table 1.

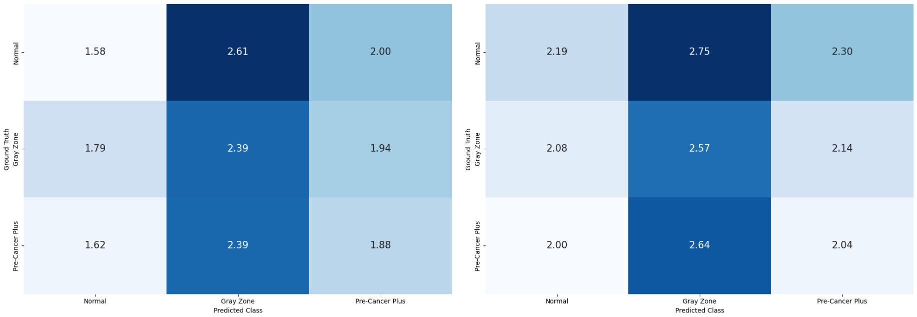

Fig. 1: Confusion Matrix of Average Conformal Prediction Length for LAC (Left) and APS (Right) with α=0.1

Table 1: T-Test for Average Prediction Length μ Comparison for Correct, Incorrect, Single, and Two-Class Misclassification (95% Confidence)

μ_1 vs μ_2	μ_1 LAC vs μ_2 LAC	μ_1 APS vs μ_2 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	< 0.05	< 0.05
Corr vs SC	1.78 ± 0.05 vs 2.43 ± 0.06	2.26 ± 0.04 vs 2.60 ± 0.05	< 0.05	< 0.05
Corr vs TC	1.78 ± 0.05 vs 1.93 ± 0.13	2.26 ± 0.04 vs 2.24 ± 0.14	< 0.05	0.81
SC vs TC	2.43 ± 0.06 vs 1.93 ± 0.13	2.60 ± 0.05 vs 2.24 ± 0.14	< 0.05	< 0.05

From Figure 1, we see differences between the various ground-truth and prediction combinations in their average conformal prediction lengths, the significance of which is tested in Table 1. After running several independent t-tests, we see that the means of the various subsets are likely not drawn from the same distribution, indicating a genuinely different treatment of these subsets by the CUQ algorithm. This was the expectation, that when the model is wrong, it is also uncertain, and one of the key findings that if it were not true, would frustrate attempts to use uncertainty clinically. Further, we see that the single-class misclassifications have larger average conformal prediction set lengths than the two-class misclassifications, an initial curiosity. 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 single-class misclassifications have larger prediction set sizes.

3.2 Relationship between MC inference uncertainty and CUQ

Figure 2 shows a box-and-whisker plot (with the calculated correlation coefficient) between conformal prediction set length and coefficient of variation of the expected values of the MC inferences and a distribution of the coefficients of variation color-code it by conformal prediction length.

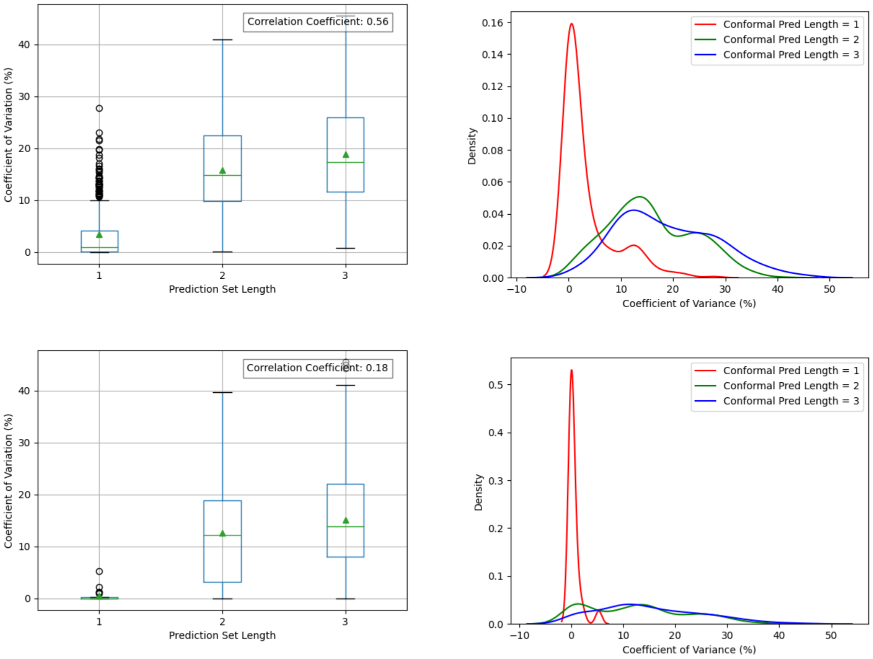

Fig. 2: Box-and-Whisker Plot of Conformal Prediction Length vs Coefficient of Variation (Left) and Distribution of Coefficient of Variation Color-Coded by Conformal Prediction Set Length (Right) for LAC (Top) and APS (Bottom) with $\alpha = 0.1$

From Figure 2, we demonstrate that as the conformal prediction set length increases, so does the coefficients of variation, and that there is a correlation between the two in the LAC case. Further, for smaller conformal prediction set lengths, the range of coefficients of variation is small and increases along with the conformal prediction set lengths. 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.

3.3 Conformal prediction set lengths by ground truth

Table 2 compares the average conformal prediction set length of the three classes based on ground truth. 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 Comparison by Ground-Truth Class (95% Confidence)

μ_1 vs μ_2	μ_1 LAC	μ_2 LAC	μ_1 APS	μ_2 APS	p LAC	p APS
Normal vs GZ	1.89 ± 0.05	2.24 ± 0.07	2.35 ± 0.04	2.43 ± 0.06	< 0.05	< 0.05
Normal vs PC+	1.89 ± 0.05	2.13 ± 0.11	2.35 ± 0.04	2.34 ± 0.09	< 0.05	0.82
GZ vs PC+	2.24 ± 0.07	2.13 ± 0.11	2.43 ± 0.06	2.34 ± 0.09	0.07	0.10

3.4 Comparing the Gray Zone image conformal prediction lengths in the two-class model

Given the existence of the “Gray Zone” class as a place to put images which do not fit neatly into “Normal” or “Precancer+”, we expect that conformal prediction applied to the two-class model will show that these images are more uncertain than their “Normal” counterparts. 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.

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 for space considerations.

Table 3: T-Test for Average Conformal Prediction Lengths μ by LAC with $\alpha = 0.1$ in the Two-Class Model (95% Confidence)

Comparison (μ_1 vs μ_2)	μ_1	μ_2	p
GZ vs Overall Inc GZ	1.51 ± 0.06	1.31 ± 0.03	< 0.05
GZ vs Overall Exc GZ	1.51 ± 0.06	1.26 ± 0.03	< 0.05
GZ vs Normal	1.51 ± 0.06	1.23 ± 0.03	< 0.05
GZ vs PC+	1.51 ± 0.06	1.46 ± 0.09	0.28

4 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 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. We expected that the model would be more uncertain of those in the two and three-class model, and for the LAC algorithm, this holds.

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, taking more 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. Prediction set lengths provide another output for the clinician to determine the next steps beyond just the model's classification, potentially aiding in edge cases or when the clinician is less experienced. 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. However, the drawback to this method is that, as this work has shown, the choice of alpha, type of CUQ algorithm, creation of the calibration set, etc., can have a marked effect on the outcomes.

5 Reproducibility

Anonymous GitHub repo

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.