Research Proposal on Enhancing Cervical Cancer Detection using AI - Addressing Data Availability and Computation Challenges

Introduction:

Every-year, more than 300,000 lives are lost to cervical cancer and, 90% of those fatalities happen in the developing countries around the world(WHO, 2021). On the one hand, the nations of the world have committed to a pledge to eliminate cervical cancer as a public health issue by 2030 under the global healthcare strategy of the WHO(WHO, 2023). On the other hand, deaths due to cervical cancer is expected to increase by 27% in the same period (Zhao et al., 2021). Developing nations around the world are fighting a heavily under-resourced battle in creating effective and efficient healthcare systems. They all lack proper infrastructure setup, financial capacity and effective governance of health care system(APHA, 2008). The state of things gets worse due to the increasing migration of qualified citizens, such as doctors and nurses, into developed nations.

The only way to win against this fatal and incurable ailment, cervical cancer, is through effective preventive measures and early-detection(NIH, 2023). Developed nations have been able to, significantly, reduce the mortality and incidence of cervical cancer (ECIS, 2021; Rahman et al., 2013). In EU-27 Nations, the cervical cancer cases accounted to 2.5% of all new incidences of cancer with even less mortality (of 2.4%) (ECIS, 2021). In Nepal, a developing nation, it is the most common cancer in women accounting for the mortality to incidence ratio of 66.5%. The overall screening coverage for cervical cancer is less than 5% in this country (ICO/IARC, 2023). The state of other developing countries is similar, if not worse. Once a globally leading cause of death in females, this illness has now become a staggeringly painful burden for the economically struggling nations. Without serious intervention, the fate of developing countries will be in a dreadful state in the coming decade.

Problem statement:

Technologies of machine-learning and deep-learning have proven to be highly efficient in diagnosing non-curable diseases such as cancers(Mangla et al., 2023) and mental-illnesses(Yadav et al., 2023). Multilayered convolution neural networks have proven to be 98.3% accurate in the diagnosis of cervical cancer, and the results were 99% precise(Zhang et al., 2020). The screening for cervical cancer comprises of two steps of tests namely pap-smear test and HPV-DNA test. This screening is followed by imagery analysis of colposcopy-images and biopsy-tests to confirm the diagnosis (Cheung et al., 2020). Each ML based model comes with its own set of limitations of biasness, interpretability, overfitting and vulnerability against given data set. Furthermore, although reliable-and-accurate, deep-learning techniques have their own trade-offs for training time, computational complexity and resource intensity(Ahmed et al., 2023). Above-all, the performance of AI based technologies is largely determined by the quality and quantity of the training data(Jiang et al., 2023).

Healthcare systems in developing countries run perennially in resource-limited settings. At such a scenario, improving the quality of existing data is a huge challenge(López et al., 2022). Training the AI model on the contextual data from these places is another serious issue(Gupta et al., 2023). Combining that to the issues of privacy, security, consent, various ethical and socio-cultural considerations, this already seems to be a lost cause for the developing nations.

The Motivation:

The battle against cervical cancer could be turned around if we can create just the right balance of accuracy and precision against the computational complexity for our predictive models. This could be achieved through a hybrid model of machine- and deep-learning techniques. The model will be an ensemble of extreme gradient boost (XGboost) and a lightweight neural-network, preferably mobilenet or squeezenet. The capability of XGboost to work with structured and engineered-dataset and its highly interpretable architecture would be an ideal-fit for a resource constrained environment. Combining that with the fast-inferencing and learning capability of mobilenet, we could get the right balance of accuracy-and-affordability with a potential for transfer-learning.

Objective:

To propose replicable and well-enhanced cervical cancer detection-system while addressing the data-issues in the healthcare sector of developing and underdeveloped-countries.

Research-questions:

1. How can we use an ensemble of XGboost with mobilenet (or squeezenet) to create a more practical model for implementing in healthcare settings with limited resources?
2. What extent of learning-transferability can be achieved if such an ensemble-model were to be developed? What extent of trade-off could be achieved for accuracy, precision, replicability and affordability?

Methodology:

Data-Extraction and Processing: The model will be trained with 3 datasets taken from a secondary data source, kaggle. First two datasets will contain images from pap-smear tests and colposcopy(CV2, n.d.; Jocelyn Dumlao, n.d.). The third, a text dataset, will be from the UCI repository, the primary-dataset around which the entire literature review of this research is based on(Gokagglers, n.d.). The image data will undergo normalization and augmentation while the text-data will be tokenized and pre-processed.

Model-Development: Image data will be preprocessed and used to train individual network of mobileNet, a lightweight learning-machine. Text data from the repository will be used to train XG Boost model. Feature engineering for both textual and image data will be implemented for improvement of XGBoost. Possibilities of learning-transfer from other pretrained-models will be pursued for faster convergence of the deep-learning system.

Ensemble-Creation: Simple averaging or stacking techniques would be used to combine the results of both models. Effective weight-adjustments will enable us to get the optimal combination. A subset, to simulate a data-sparse-scenario, that will be created from training datasets above, will be the test-data.

Model-Optimization: XGboost will further undergo hyperparameter-tuning for efficiency. Fine-tuning of neural-networks for model compression and quantization to ensure easy-deployment. Evaluation of interpretability, computational-overhead, cost-effectiveness and overall-efficiency for resource-constrained setting.

Evaluation-and-generalization: The results will be measured for accuracy, sensitivity, F1-score, precision and area-under-the-curve. The robustness of this model, if delivered, could be accessed on external datasets to validate its potential for broader-application.

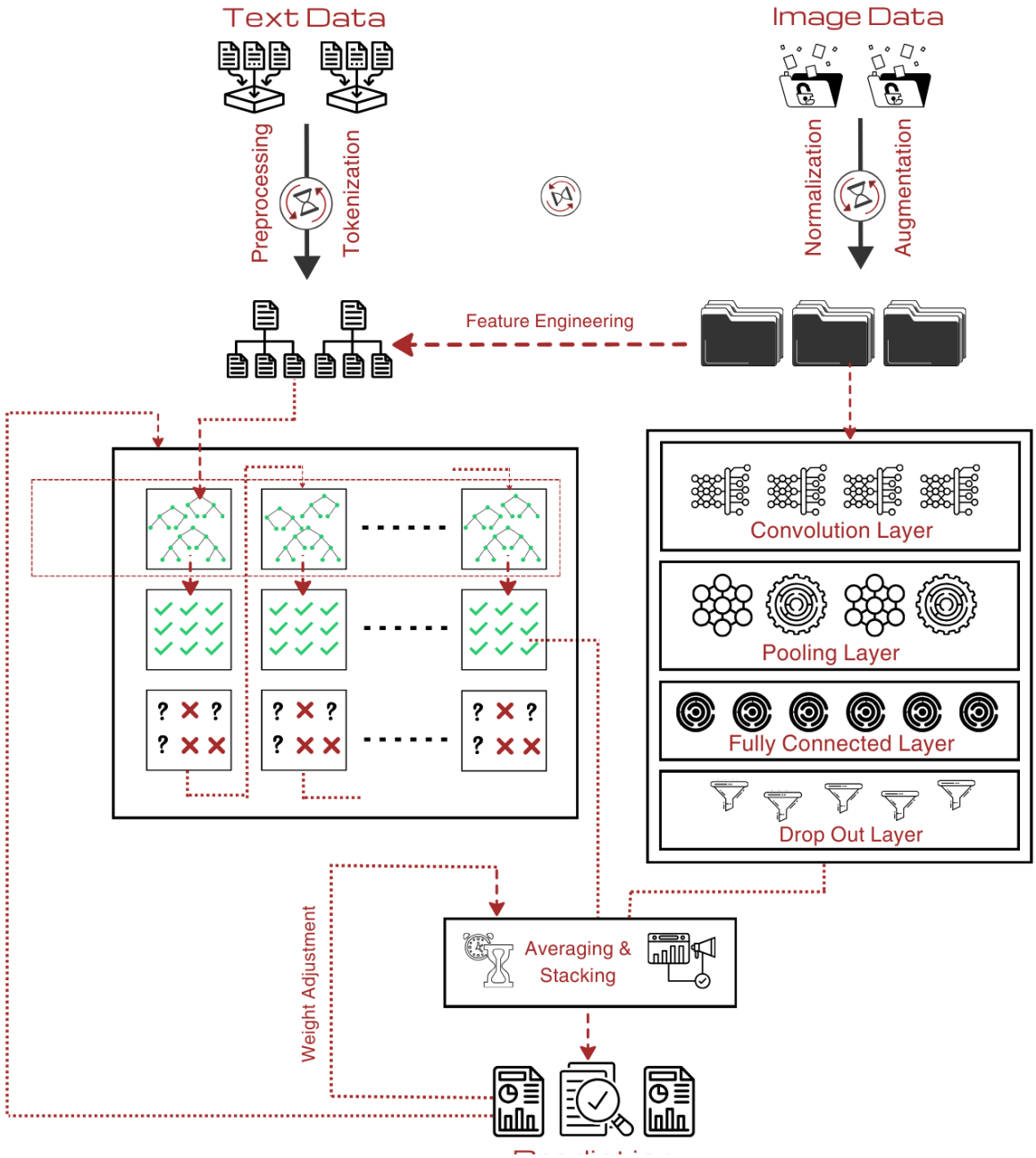


Figure 1 Architecture of the Proposed System

Limitations:

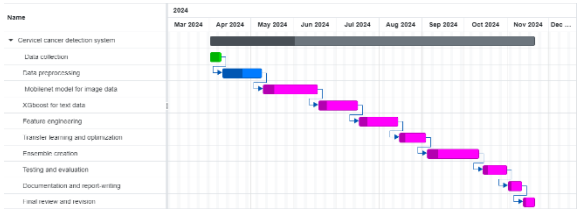
The design philosophy of this model is solely based on the parameters of a resource-limited-setting. Hence, a few limitations of this model may arise, such as:

- Limited capability of XGboost to handle complex data and their inter-relationship. The output highly depends on weight-adjustments for decision of each model.
- Hampered learning transferability, mostly due to varying demographics and healthcare practices across developing nations.
- Subset that mimics sparse-data-scenario in this test may or may not accurately represent the real-world scenario.

Ethical-considerations:

Although the proposed model uses secondary datasets for testing purposes, this model will follow all the ethical principles mandated by the governing curriculum to ensure best practices. The implementation of this model will subsequently follow best ethical practices including transparency, informed-consent, fairness and collaboration with all stakeholders involved.

Gantt-Chart:



Conclusion:

This research aims to develop a hybrid of machine-learning and deep-learning techniques to create just the right fit for resource-limited healthcare environment. A hybrid model, that works on both text-and-image data, that can learn fast and in a lean-mode, that works-well for sparse-datasets. It can (efficiently) learn from other pretrained models without becoming too resource-intensive. And the learning from one model could be seamlessly transferable to other-similar-models. If such a model could be developed, this could, single-handedly, change way developing nations are combating the battle against cervical cancer.

References:

Ahmed, S. F., Alam, M. S. Bin, Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M., & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, *56*(11). https://doi.org/10.1007/s10462-023-10466-8

APHA. (2008). Strengthening Health Systems in Developing Countries. *Apha.Org*.

Cheung, L. C., Egemen, D., Chen, X., Katki, H. A., Demarco, M., Wiser, A. L., Perkins, R. B., Guido, R. S., Wentzensen, N., & Schiffman, M. (2020). 2019 ASCCP Risk-Based Management Consensus Guidelines: Methods for Risk Estimation, Recommended Management, and Validation. *Journal of Lower Genital Tract Disease*, *24*(2), 90–101. https://doi.org/10.1097/LGT.0000000000000528

CV2. (n.d.). *Cervical Cancer Image Dataset (SipkaMed)*. Kaggle. Retrieved March 5, 2024, from https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed

ECIS. (2021). *Cervical cancer burden in EU-27*. https://cancer-code-europe.iarc.fr

Gokagglers. (n.d.). *UCI ML Repository*. Kaggle. Retrieved March 5, 2024, from https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification

Gupta, R., Kumar, N., Bansal, S., Singh, S., Sood, N., & Gupta, S. (2023). Artificial Intelligence-driven Digital Cytology-based Cervical Cancer Screening: Is the Time Ripe to Adopt This Disruptive Technology in Resource-constrained Settings? A Literature Review. In *Journal of Digital Imaging* (Vol. 36, Issue 4). https://doi.org/10.1007/s10278-023-00821-0

ICO/IARC. (2023). *Nepal:Human Papillomavirus and Related Cancers, Fact Sheet 2023*. www.hpvcentre.net

Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J., & Liu, J. (2023). A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artificial Intelligence Review*, *56*, 2687–2758. https://doi.org/10.1007/s10462-023-10588-z

Jocelyn Dumlao. (n.d.). *Malhari Dataset*. Kaggle. Retrieved March 5, 2024, from https://www.kaggle.com/datasets/jocelyndumlao/malhari-dataset

López, D. M., Rico-Olarte, C., Blobel, B., & Hullin, C. (2022). Challenges and solutions for transforming health ecosystems in low- and middle-income countries through artificial intelligence. In *Frontiers in Medicine* (Vol. 9). https://doi.org/10.3389/fmed.2022.958097

Mangla, S., Saini, P., Jayswal, A. K., Sanyal, K., & Pal, S. (2023). An Ai Based Application for Cancer Diagnosis - An Emperical Analysis. *Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2023*, 231–236. https://doi.org/10.1109/Confluence56041.2023.10048847

NIH. (2023). Cervical Cancer Prognosis and Survival Rates. *Cervical Cancer*. https://www.cancer.gov/types/cervical/survival

Rahman, M., Mia, A. R., Haque, S. E., Golam, M., Purabi, N. S., & Choudhury, S. A. R. (2013). *Current Topics in Public Health* (Alfonso J. Rodriguez-Morales, Ed.).

WHO. (2021). *Cervical Cancer* . https://www.who.int/health-topics/cervical-cancer#tab=tab_1

WHO. (2023). *Cervical Cancer Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

Yadav, A., Kumar, D., & Hasija, Y. (2023). Behaviour Analysis Using Machine Learning Algorithms in Health Care Sector. *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023*, 877–880. https://doi.org/10.1109/InCACCT57535.2023.10141829

Zhang, Y., Li, L., Gu, J., Wen, T., & Xu, Q. (2020). Cervical Precancerous Lesion Detection Based on Deep Learning of Colposcopy Images. *Journal of Medical Imaging and Health Informatics*, *10*(5). https://doi.org/10.1166/jmihi.2020.3051

Zhao, M., Wu, Q., Hao, Y., Hu, J., Gao, Y., Zhou, S., & Han, L. (2021). Global, regional, and national burden of cervical cancer for 195 countries and territories, 2007–2017: findings from the Global Burden of Disease Study 2017. *BMC Women's Health*, *21*(1). https://doi.org/10.1186/s12905-021-01571-3