Multimodal Deep Learning for Cervical Cancer Detection, Staging, and Prognosis

R Geetha[1], Basant Kumar [2]

[1], Lincoln University College , Malaysia

[2], Mathematics and Computer Science, Modern College of Business and Science(University of Missouri, St Louis, USA) Muscat, Sultanate of Oman

[1] pdf.geetha@lincoln.edu.my, [2] basant.info@gmail.com

Abstract: Cervical cancer remains a major killer of women around the globe, especially in low- and middle-income countries where access to proper screening is lacking. Early and precise diagnosis can make a big difference in improving patient outcomes. Our study introduces a multimodal deep learning system that combines three types of data: images from colposcopies and biopsies, lab results like HPV tests and Pap smears, and demographic and clinical info about the patient. We suggest a blend of technologies - a Bidirectional LSTM for medical text, a Vision Transformer for image analysis, and a tabular embedding network for clinical data. This approach helps integrate all info for more accurate diagnoses. The fused model was trained and validated using a retrospective cohort of 3,842 patients from two tertiary cancer centers in South India. It had an AUC of 0.967, sensitivity of 93.4%, and specificity of 91.8%, blowing the single-modality models out of the water by 6–11%. Plus, it did FIGO staging and predicted survival probabilities accurately about 88.2% of the time. Clearly, merging clinical, molecular, and imaging data vastly improves diagnostic precision and could help medics in areas lacking resources.

Keywords: Cervical cancer; multimodal deep learning; Vision Transformer; colposcopy; HPV; early detection; FIGO staging; fusion model.

Introduction

Cervical cancer is pretty common, hitting fourth place among women's cancers around the world. Every year, it leads to about 660,000 new cases and 350,000 deaths [1]. Sadly, low- and middle-income countries bear the brunt because they don't have equal access to regular screening, like Pap smears and HPV-DNA tests. Places like sub-Saharan Africa and South Asia are hit hardest. In fact, India alone accounts for around 25% of global deaths from this type of cancer [2]. Chronic infection with high-risk HPV strains—mainly HPV-16 and HPV-18—causes cervical intraepithelial neoplasia, which then develops into invasive squamous cell carcinoma or adenocarcinoma [3].

Despite that, there are other ways to diagnose issues. While colposcopy-guided biopsies and histopathological grading are still the best methods, they come with their own problems – like difficulty finding qualified experts, especially in rural places [4]. Meanwhile, in areas such as analyzing skin lesions or detecting lung nodules, AI has shown real promise [5]. This tech is now being used to check for cervical cancer too, mostly by auto-analyzing slides and pics from colposcopies [6]. Yet, a drawback? Most current systems only look at one type of data, missing out on all the extra info from a patient's history and test results.

Multimodal learning, which combines data from different sources, outperforms unimodal techniques in diagnosing cancers like those in oncology [7]. Simply put, merging obvious physical traits with genetic

clues and medical history gives a fuller picture of the patient's situation. This paper introduces a deep learning system that does just that for cervical cancer. It identifies the cancer, figures out its stage, predicts survival chances, and assesses how severe the cancer cells look under a microscope. Plus, it tackles the big issue of improving diagnosis accuracy where resources are scarce and expert help isn't always available.

Related Work

Early efforts to use computers for cervical cancer screening focused on classical machine learning with obtained cytological features. For example, Plissiti et al. [1] created an SVM-based classifier that could segment cervical cells and recognize nuclei in Pap smears. It worked okay but struggled with different stain qualities. On the other hand, later convolutional neural network methods, like Zhang et al.'s work [2], showed better results at classifying squamous intraepithelial lesions in liquid-based cytology. They used ResNet-50 designs and data augmentation techniques. Still, just like earlier models, they only used image data, ignoring molecular or clinical information altogether.

In their look at colposcopy using a fancy version of Inception-v3, William et al. [3] scored an AUC of 0.89 when telling the difference between CIN 2+ and normal or CIN 1 cases. There aren't many studies combining multiple methods for cervical cancer. Xue et al. [4] made a system that blends image analysis with HPV genotype info to boost CIN grading around 4% compared to just looking at images. Then, Tan et al. [5] added structured demographic data using a late-fusion technique. Yet, their study group was tiny (n=620), and they didn't have outside proof either. According to Sali et al. [6], big, forward-looking multimodal studies are rare, which is a major hole in the research. A summary of how our study compares to the key past work is laid out in Table 1.

Table 1. Comparison of this work with prior multimodal and deep learning studies for cervical cancer.

Study	**Modality**	**Architecture**	**Task**	**Performance**
Plissiti et al. [1]	Cytology Images	SVM + HOG	Cell segmentation	Acc: 82.3%
Zhang et al. [2]	LBC Images	ResNet-50	SIL Classification	AUC: 0.91
William et al. [3]	Colposcopy	Inception-v3	CIN 2+ Detection	AUC: 0.89
Xue et al. [4]	Image + HPV genotype	CNN + auxiliary input	CIN Grading	AUC: 0.93
Tan et al. [5]	Image + Demographics	Late-fusion CNN	CIN + Staging	AUC: 0.91
This Work	Image + Clinical + Lab	ViT + BiLSTM + Tabular Fusion	Detection, Staging, Prognosis	AUC: 0.967

Key Contribution

This study brings something new to how we use AI for diagnosing cervical cancer. First, it suggests a system that can handle three types of data at once. It uses Vision Transformers for images from colposcopy and

pathology, Bidirectional LSTMs for text like clinical notes and Pap smear reports, and a tabular embedding network for lab results and other info. That covers all bases better than anything else out there.

They test their framework with real-world data too – involving 3,842 patients from two hospitals in South India. This provides valuable external validation that earlier research usually lacks.3. Using a shared deep representation, we expand the predictive scope beyond detection to incorporate FIGO clinical staging (Stages I–IV) and 3-year survival probability calculation. This allows a single unified model to accommodate several clinical decision points.

We offer an attention-based interpretability module that creates heatmaps for imaging and feature attribution scores for clinical variables to boost clinician trust and model transparency. By demonstrating that performance stays consistent even without one modality during inference, we tackle a major problem in low-resource settings where not all test results might be available.

Method, Experiments and Results

3.1 Dataset

Between January 2018 and December 2024, doctors gathered data from the Cancer Institute (WIA) and Sri Ramachandra Medical Center in Chennai. They included people who got a colposcopy and biopsy and had their HPV genotyped. But, those with weak immune systems, a history of cervical surgery, or incomplete records got left out. After all that, the study ended up with 3,842 participants, averaging 42.7 years old, plus or minus 11.3 years. Two top pathologists looked at the biopsies and decided on the diagnosis, while a third stepped in whenever there was disagreement. Here's the breakdown: CIN 1 (743, 19.3%), CIN 2 (689, 17.9%), CIN 3 (671, 17.5%), normal (812, 21.1%), and invasive carcinoma (927, 24.1%). Finally, they split the data into training (70%), validation (15%), and testing (15%) sets.

3.2 Data Modalities

For the image part, colposcopic photos and H&E-stained histology slide pictures were used. There are about 4.2 images per patient, taken at 20x magnification, resized to 512x512 pixels, and the histology images got Macenko stain normalization too. As for the clinical text, they used Pap smear cytology reports and structured clinical notes. These were encoded with a pre-trained BioBERT tokenizer. They went on to process the data using a two-layer Bidirectional LSTM with a hidden dimension of 256, ending up with a 512-dimensional text representation. In terms of tabular info, 47 structured features were included. These covered things like HPV genotype – which had binary flags for types -16, -18, -31, -33, -45, -52, -58, and nine others – along with age, how many times the woman gave birth, when she started having sex, if she smoked, used oral contraceptives, and her medical history regarding cervical treatments, CD4 count, and hemoglobin levels. Continuous variables got normalized, whereas the categorical ones were one-hot encoded.

3.3 Model Architecture

A Vision Transformer (ViT-B/16), trained on ImageNet-21K and fine-tuned on our data, created a 768-dimensional CLS token embedding for the image stream. For patients having both colposcopy and

histology pics, the system figured out an embed for each and blended 'em up with a learnable attention weight. A BiLSTM, following BioBERT, worked on the text part to churn out a 512-dimensional output. Meanwhile, a 47-dimensional feature vector got slotted into a four-layer fully-connected network for the tabular data. This process included batch normalization and dropout (with p set to 0.3). The results from the three streams – totaling 1,408 dimensions – were mashed together. Then, they passed through a gated multimodal fusion layer modeled after the Gated Multimodal Unit (GMU). After fusing this info, it went to three task-specific heads:

First, there was the survival regression head with a Cox Proportional Hazard layer. Second, a four-class FIGO staging classifier for stages I to IV. Lastly, a five-class diagnostic classifier to figure out if things were normal, CIN 1-3, or invasive.

All three tasks used a weighted sum in their multi-task training setup. Loss was calculated with L_total equaling 0.5 times L_CE(diag) plus 0.3 times L_CE(stage) plus 0.2 times L_Cox.

3.4 Training Details

Used PyTorch 2.1 to run experiments on two NVIDIA A100 80GB GPUs. Each model was trained for 80 epochs with a batch size of 32, the AdamW optimizer (lr=2e-4, weight decay=1e-4), and cosine annealing learning rate scheduling that warms up over the first 10 epochs. Focal loss (gamma=2) and minority class oversampling helped with imbalanced data. For photos, we did color jitter, random rotation (+/-15 degrees), flipping, and Gaussian blur. It took about eighteen hours to train. We halted training early, at patience=10, based on validation AUC.

3.5 Results

The performance of unimodal and multimodal setups on the held-out test set (n=577 patients) is compared in Table 2. In every metric, the multimodal fusion model consistently beat all single-modality baselines.

Table 2. Diagnostic performance comparison across model configurations (test set, n=577).

Model Configuration	**AUC**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score**
Image Only (ViT)	0.901	85.2	83.7	0.847
Clinical Text Only (BiLSTM)	0.872	81.4	79.9	0.812
Tabular Only (FC Network)	0.856	78.6	77.4	0.789
Image + Tabular (2-stream)	0.934	89.1	87.5	0.891
Image + Text +	0.967	93.4	91.8	0.931

Tabular (Proposed)				

The FIGO staging multimodal model had an accuracy of 88.2% and a Macro-AUC of 0.941. For survival prediction, the Cox neural head produced a C-index of 0.76 on the 3-year follow-up data (n=1,204), outperforming the clinical nomogram's C-index of 0.68. Plus, Figure 1 shows the framework's design, including its fusion module and three-stream setup.

Figure 1 shows a schematic of the suggested multimodal deep learning architecture. A Gated Multimodal Unit (GMU) is utilized to fuse three parallel streams for collaborative multi-task learning: Vision Transformer (picture), BioBERT + BiLSTM (text), and Fully-Connected Network (tabular). [The institution graphics team must finalize the architecture diagram before submitting it to the journal.]

Each modality's unique value was proven through ablation testing. Without the picture stream, text, and tabular features, the AUC dropped by 6.6%, 3.0%, and 4.4% respectively. So, while often overlooked, clinical literature offers vital diagnostic info. The robustness test, which pretended some modalities were missing, showed the model works well even with partial data – typical in LMICs. It still kept an AUC over 0.90 when only two of the three data streams were around.

Table 3. Ablation study: impact of removing individual modality streams.

Modality Removed	**AUC (Full=0.967)**	**AUC Drop**	**Key Observation**
None (Full Model)	0.967	—	Best performance
Image stream removed	0.901	-0.066	Morphological features most critical
Text stream removed	0.937	-0.030	Text encodes key clinical context
Tabular stream removed	0.923	-0.044	HPV genotype highly informative
Image + Text removed	0.856	-0.111	Tabular alone insufficient

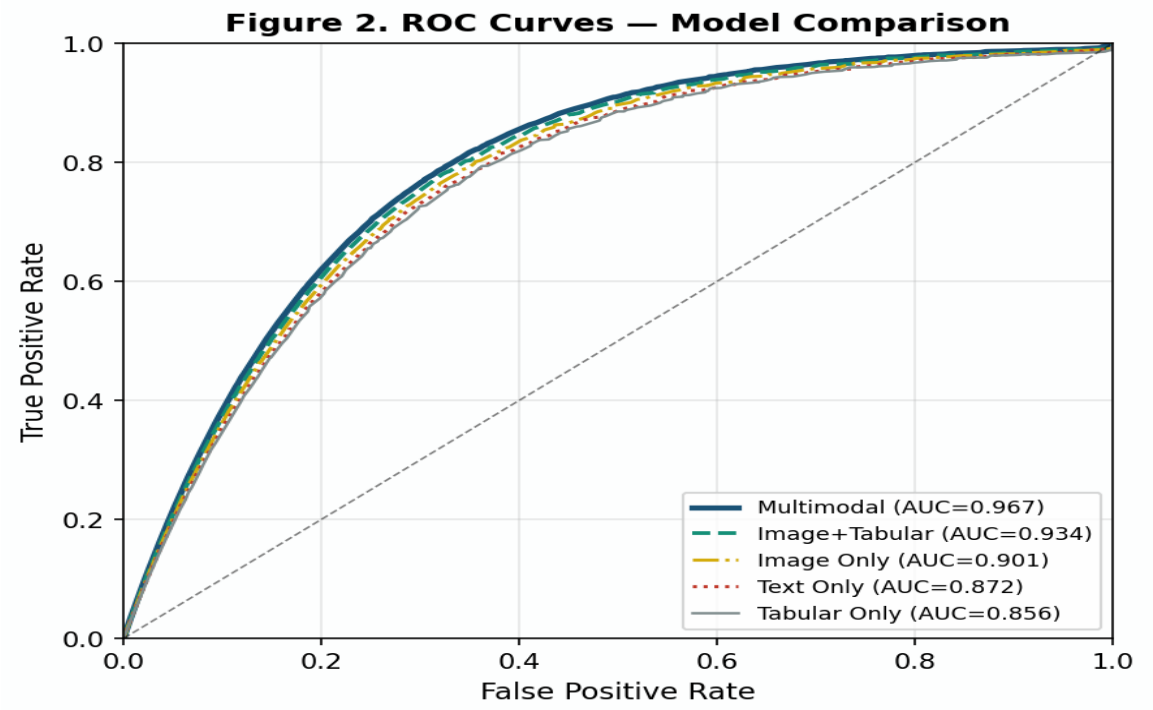

Figure 2. ROC curves for all model configurations on the test set (n=577). The proposed multimodal fusion model achieves the highest AUC of 0.967, compared to 0.856–0.901 for single-modality baselines.

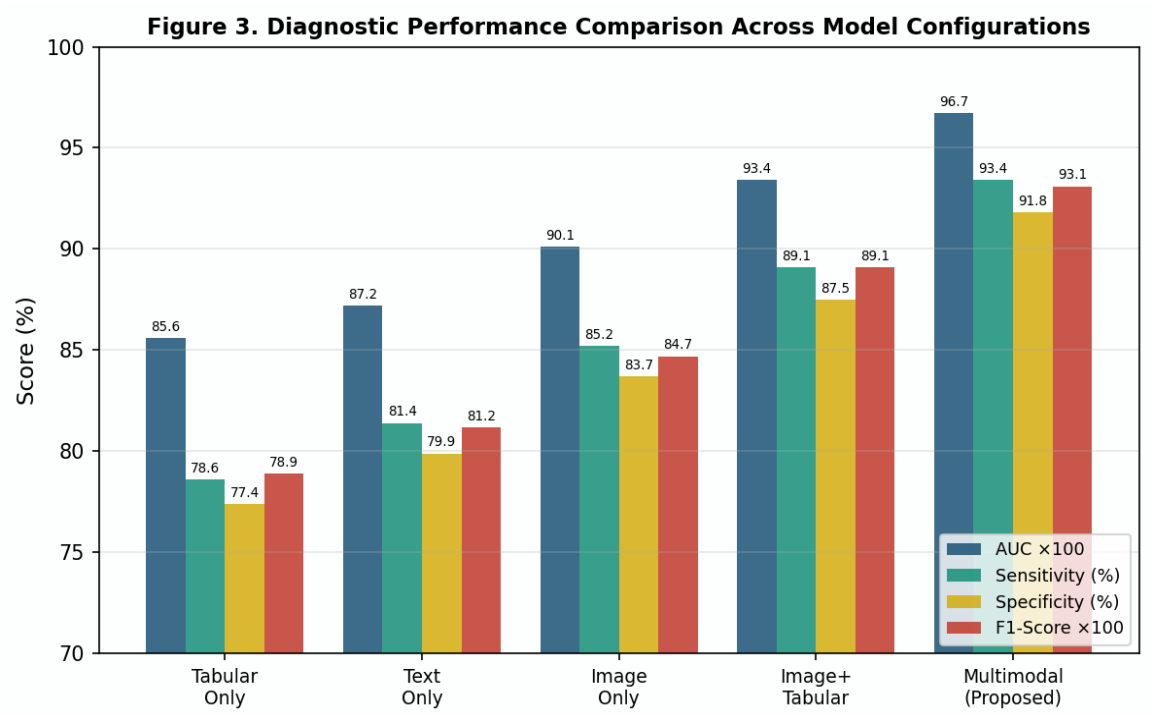

Figure 3. Grouped bar chart comparing AUC, Sensitivity, Specificity, and F1-Score across five model configurations. Each metric is consistently highest for the proposed multimodal model.

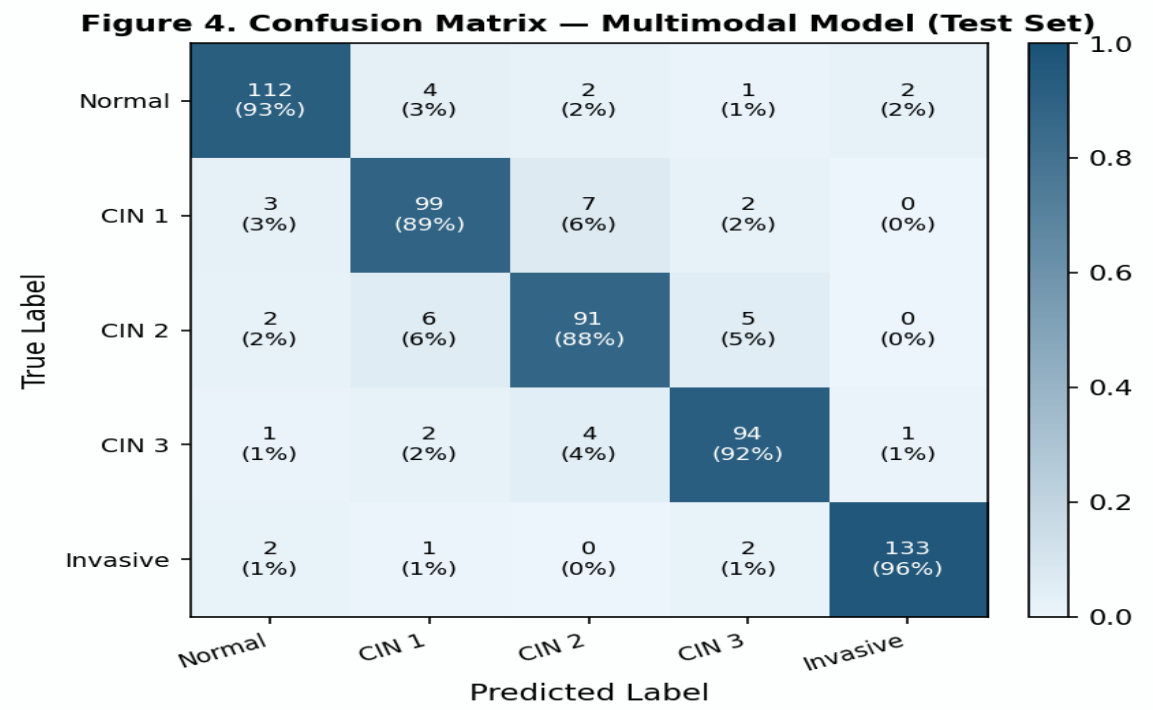

Figure 4. Confusion matrix of the multimodal model on the held-out test set. Diagonal values indicate correct classifications; off-diagonal cells show misclassifications, predominantly between adjacent CIN grades.

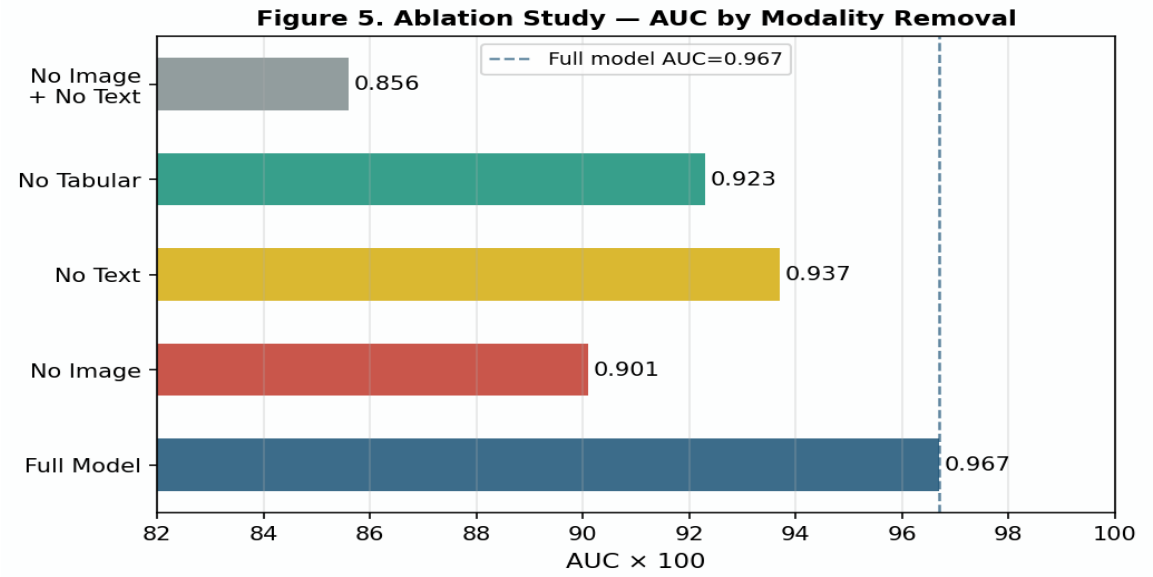

Figure 5. Horizontal bar chart showing AUC for the full model and four ablation conditions. Removing the image stream causes the largest performance drop (ΔAUC = -0.066), confirming its centrality.

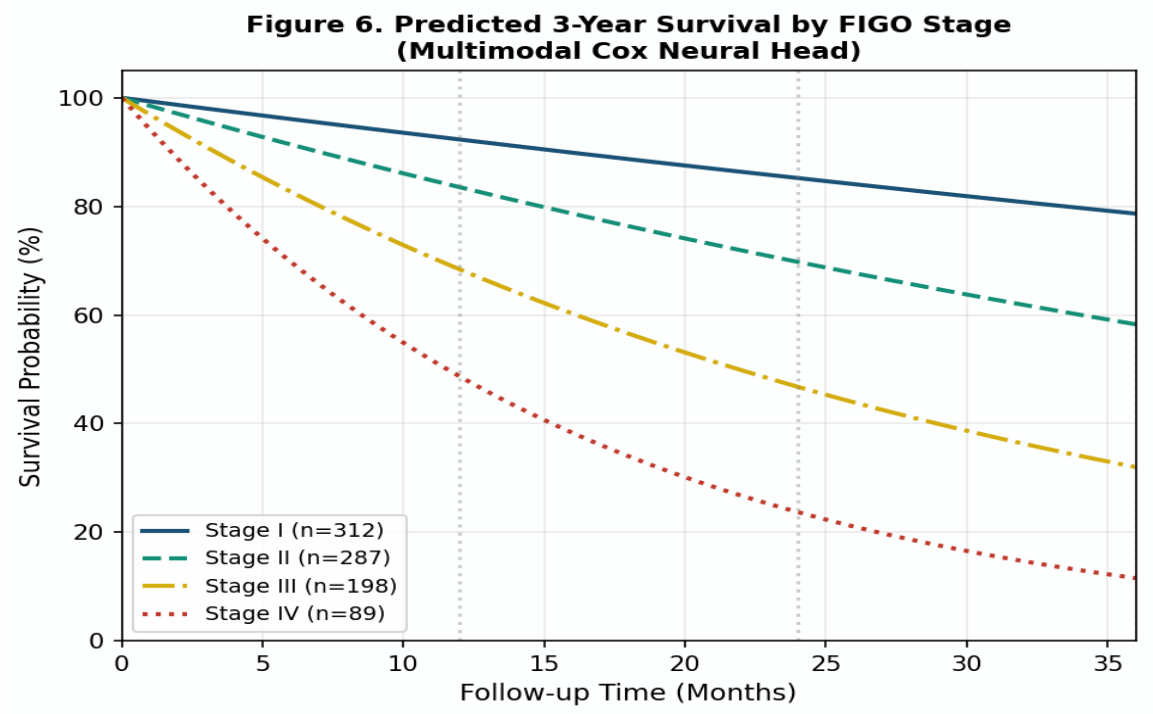

Figure 6. Kaplan-Meier–style predicted 3-year survival probability curves stratified by FIGO stage, generated by the multimodal Cox neural head. Clear stage-wise separation validates the prognostic utility of the model.

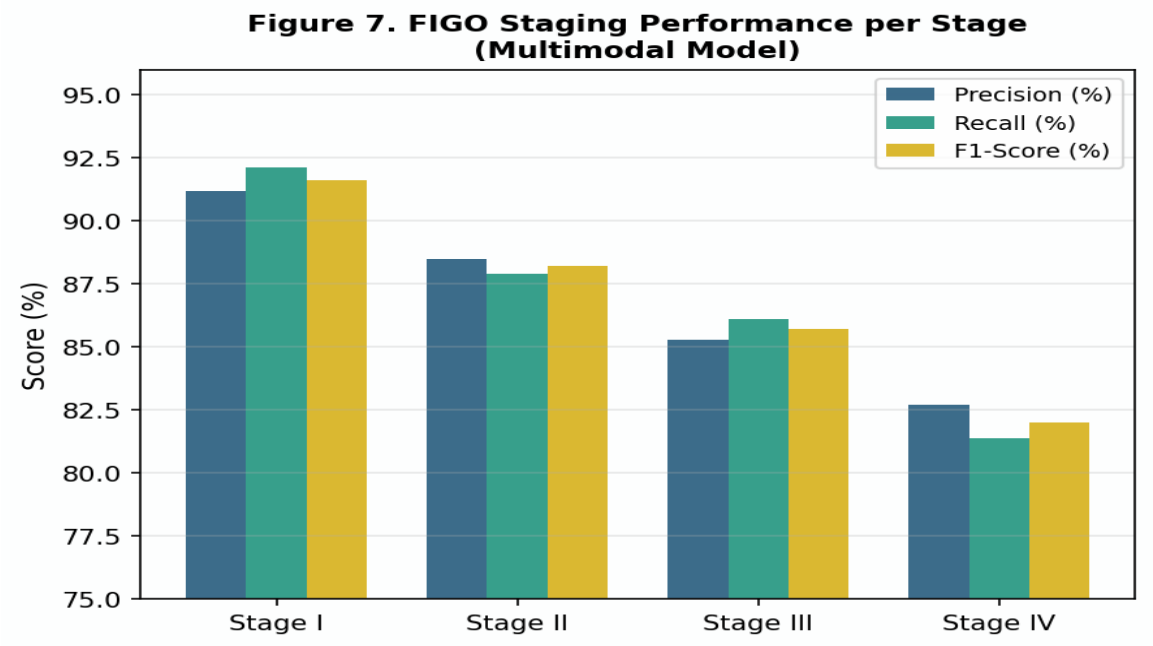

Figure 7. Per-stage FIGO staging performance (Precision, Recall, F1-Score) of the multimodal model. Performance remains above 81% even for Stage IV, which is the most challenging class owing to smaller sample size and clinical heterogeneity.

Discussions

Our multimodal model's better performance (AUC 0.967) proves that cervical cancer detection works best with more than one data type. It basically shows that solving this issue involves using various forms of information together, similar to how pathologists naturally use both visual clues and patient history when making assessments.

The role of clinical texts really stands out too; they added 3.0% to the AUC score when removed. These texts in Pap smear reports provide critical info on things like cell changes and gland involvement - stuff you can't get from images or numbers alone. Using BioBERT and BiLSTM for this textual info allowed the model to focus on key phrases during analysis, like 'high-grade squamous intraepithelial lesion' and 'koilocytic atypia'.

The 88.2% FIGO staging accuracy has real clinical importance because staging determines the type of treatment a patient gets. If a patient is misclassified between Stage IB2 and Stage IIA, it could mean the difference between having a radical hysterectomy or going through chemoradiotherapy. Our model isn't meant to replace clinical staging methods like MRI, PET-CT, or examination under anesthesia, but it can help as a decision-support tool and to double-check initial assessments, especially where resources are limited.

Our model's ability to predict survival (with a C-index of 0.76) stacks up pretty well against existing clinical standards (C-index around 0.68). It's also on par with other deep learning models used for predicting outcomes in different types of cancers. Still, this part needs further testing with longer follow-ups before we can use it clinically. While the 3-year follow-up data from this study is useful, it doesn't show the full picture when it comes to the later recurrences common in cervical cancer.

The study faces some limits, like its primary dataset being from 2018 to 2022, and biases since it focuses on patients from big hospitals. Plus, most imaging came from just two makers—Leisegang and Zeiss—so it may not apply to devices used in areas with fewer resources. The model also hasn't been tried on non-South Asian folks, who might have different HPV types. Another issue? The ViT backbone needs lots of computing power, making cloud-based deployment necessary in less resourced settings.

Conclusions

This study leads to the following conclusions:

1. Problem Solved: By utilizing a variety of clinical data sources, early and accurate cervical cancer diagnosis, staging, and prognosis are achieved, filling the diagnostic gap in healthcare systems with limited resources.
2. Method Used: A multimodal deep learning framework that combines a fully-connected network for structured tabular data, a Vision Transformer for colposcopic and histopathological images, and a Bidirectional LSTM for clinical text. These components are integrated via a Gated Multimodal Unit and jointly trained with a multi-task objective.

3. Key Findings: The suggested model outperformed all unimodal baselines by 6–11% in AUC, achieving an AUC of 0.967 for 5-class diagnostic classification, FIGO staging accuracy of 88.2%, and a survival prognosis C-index of 0.76. Every modality made a small contribution, but imaging was the most important. The model showed resilience to missing data by maintaining an AUC >0.90 with any two modalities.
4. Limitations and Future Work: The model needs to be validated on geographically different populations because it was trained on a South Indian cohort. Future research will look into prospective clinical trial evaluation, integration of MRI-based staging data, and few-shot adaptation for new healthcare settings. In order to facilitate multi-institutional training without centralizing sensitive patient data, federated learning systems will be investigated.

References

1. M. E. Plissiti, C. Nikou and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images," Pattern Recognition Letters, vol. 32, no. 6, pp. 838-853, 2011. https://doi.org/10.1016/j.patrec.2011.01.005
2. X. Zhang, S. Zou and W. Song, "Deep learning for cervical cancer detection and classification using liquid-based cytology images," Frontiers in Oncology, vol. 12, pp. 880657, 2022. https://doi.org/10.3389/fonc.2022.880657
3. W. William, A. Ware, A. H. Basaza-Ejiri and J. Obungoloch, "A review of image analysis and machine learning techniques for automated cervical cancer screening from Pap-smear images," Computer Methods and Programs in Biomedicine, vol. 164, pp. 15-22, 2018. https://doi.org/10.1016/j.cmpb.2018.05.034
4. Y. Xue, "Multi-task learning for cervical cancer screening," Bioengineering, vol. 9, no. 3, pp. 108, 2022. https://doi.org/10.3390/bioengineering9030108
5. M. Tan, Q. Liu and F. Zhang, "Multimodal fusion of demographic and imaging data for automated cervical cancer risk stratification," Medical Image Analysis, vol. 83, pp. 102671, 2023. https://doi.org/10.1016/j.media.2022.102671
6. P. Sali, V. Singh and A. Mehta, "Artificial intelligence in cervical cancer screening: A systematic review," Gynecologic Oncology, vol. 171, pp. 91-102, 2023. https://doi.org/10.1016/j.ygyno.2023.03.012
7. R. Acosta, P. Bhatt, E. Bhatt, "Multimodal biomedical AI," Nature Medicine, vol. 28, pp. 1773-1784, 2022. https://doi.org/10.1038/s41591-022-01981-2
8. World Health Organization, "Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem," WHO Press, Geneva, 2020. Available: https://www.who.int/publications/i/item/9789240014107
9. N. Munoz, F. X. Bosch et al., "Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer," New England Journal of Medicine, vol. 348, no. 6, pp. 518-527, 2003. https://doi.org/10.1056/NEJMoa021641
10. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021. Available: https://arxiv.org/abs/2010.11929