

Cervix-Aware Region-Based Ensemble Graph Neural Networks for Cervical Cancer Detection from Colposcopy Images

by

Ronodeep Bhowmik

24141083

Anwesha Roy

24141100

Shaptarshi Saha

23341094

Miftahul Jannat Orpi

21101165

Monjur Ahmed

21301642

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ronodeep Bhowmik
24141083

Anwasha Roy
24141100

Shaptarshi Saha
23341094

Miftahul Jannat Orpi
21101165

Monjur Ahmed
21301642

Approval

The thesis/project titled “Cervix-Aware Region-Based Ensemble Graph Neural Networks for Cervical Cancer Detection from Colposcopy Images” submitted by

1. Ronodeep Bhowmik(24141083)
2. Anwesha Roy(24141100)
3. Shaptarshi Saha(23341094)
4. Miftahul Jannat Orpi(21101165)
5. Monjur Ahmed(21301642)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Cervical cancer continues to pose an enormous risk to global health, underscoring the urgent necessity for prompt and accurate detection to improve the health of women affected by this deadly illness. Current advancements on graph neural networks exhibit potential in classifying medical images, yet existing models usually demand massive computational loads, making their implementation unaffordable for clinically resource-limited settings. This research presents an intriguing ensemble learning framework that combines various graph neural network architectures, such as GraphMAE2 (large and small variants), GAT, GIN, and GCN, aimed at improving the classification accuracy for cervical cancer identification. Moreover, our approach utilizes the colposcopy images, acquired from the dataset, and transforms them into graph representations enriched with handcrafted medical features and cervix-specific annotations. This provides a detailed and clinically contextualized description of the cervical area. Our ensemble model utilizes an innovative training approach featuring pre-training by masked autoencoding and fine-tuning with cross-validation to arrive at peak performance and generalizability on small datasets. The proposed ensemble model exhibits enhanced performance, attaining a validation accuracy of 95.52% and an F1 score of 94.12%, thereby exceeding the results of individual models. Additionally, we incorporate uncertainty quantification for enhancing the reliability of predictions and it provides insightful feedback on the confidence of predictions. This makes our framework an informative guidance tool for cervical cancer screening, with specific utility for clinically resource-limited settings. This contribution enhances the expanding domain of graph-based processing of medical images and delivers a robust and generalizable solution for healthcare systems.

Keywords: Data Imbalance, Supersixel, Feature masking, Self-supervised learning, GCN, GAT, GraphMAE2, GIN, Class weight, Uncertainty.

Acknowledgement

First and foremost, we are grateful to the Great Allah for allowing us to finish our thesis without any significant setbacks. Second, thanks to Dr. Ashraful Alam Sir, our advisor, for his insightful advice and support during the course of our work. Whenever we needed help, he was there to help and guide us.

Lastly, we would want to express our gratitude to our loving parents for their unwavering support and prayers. We are now close to graduating thanks to their gracious assistance and prayers.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
List of Tables	vii
Nomenclature	viii
1 Introduction	1
1.1 Research Motivation	2
1.2 Research Problem	3
1.3 Research Objective	4
1.4 Research Contribution	4
2 Related Works And Background	6
2.1 Deep learning Model Used in Cervical Cancer Classification	6
2.2 Graph Architecture Utilized in Medical Image Diagnosis	12
3 Methodology	18
3.1 Structure of Our Proposed Model	18
3.2 Dataset Collection	19
3.3 Dataset Pre-Processing	20
3.3.1 Dataset Standardization	20
3.3.2 Dataset Splitting	20
3.3.3 Dataset Augmentation	21
3.3.4 Annotation	23
3.3.5 Feature extraction	24
3.3.6 Graph Construction	28
3.4 Selected Models Description	29
3.4.1 GraphMAE2	30
3.4.2 Graph Attention Network (GAT)	30
3.4.3 Graph Isomorphism Network (GIN)	31
3.4.4 Graph Convolutional Networks (GCN)	31

4	Model Implementation	33
4.1	Proposed GNN Architecture Components	33
4.1.1	Customized Graph Masked Autoencoder (GraphMAE2) . . .	33
4.1.2	Customized Graph Isomorphism Network (GINModel) . . .	34
4.1.3	Customized Graph Convolutional Network (GCNModel) . . .	35
4.1.4	Customized Graph Attention Network (GATModel)	36
4.2	Loss Functions and Classification Components	38
4.2.1	Focal Loss Implementation	38
4.2.2	Ensemble Classifier	39
4.3	Core Ensemble Architecture	39
4.3.1	Model Integration	39
4.3.2	Pre-Training Methodology	40
4.3.3	Forward Propagation Logic	41
4.4	Training Framework	41
4.4.1	Cross-Validation Setup	41
4.4.2	Model Training	42
4.4.3	Training Loop Implementation	43
4.4.4	Fine-Tuning Phase	43
5	Performance Evaluation	45
5.1	Experimental Setup	45
5.2	Evaluation Metrics	45
5.2.1	Accuracy	46
5.2.2	Precision	46
5.2.3	Recall	46
5.2.4	F1 Score	46
5.2.5	Uncertainty	47
5.3	Comparative Analysis	47
5.3.1	Performance Analysis of GraphMAE2 Model	47
5.3.2	Performance Analysis of the Ensemble Graph Neural Network	48
5.3.2.1	Cross-Validation Performance	48
5.3.2.2	Fine-Tuning Performance	55
5.3.2.3	Uncertainty-Threshold Performance	55
5.4	Confusion Matrix and Error Analysis	56
5.5	Key Findings and Significance	57
6	Limitations And Future Work	58
6.1	Limitations	58
6.2	Future Work	59
7	Conclusion	60
	Bibliography	65

List of Figures

3.1	Proposed Model Workflow Diagram	19
3.2	Dataset Split and HPV Distribution	21
3.3	Moderate Augmentation	23
3.4	Cervix zone annotation from colposcopy image	24
3.5	Feature Extraction from Colposcopy Images	25
3.6	Average Feature Values by Type Distribution	26
3.7	Medical Features and Feature Group by Importance	26
3.8	Comparative Analysis of Clinical Feature Distributions for HPV Status	27
3.9	PCA and t-SNE Feature Embedding and Feature Group Separability	28
3.10	Graph construction from original image	29
3.11	GraphMAE2 Model Stucture	30
3.12	Graph Attention Network model Structure [39]	31
3.13	Graph Isomorphism Network Model Structure [9]	31
3.14	Graph Convolutional Network model Structure [3]	32
4.1	Architecture of Customized GraphMAE2 Model	34
4.2	Architecture of Customized GIN Model	35
4.3	Architecture of Customized GCN Model	36
4.4	Architecture of Customized GAT Model	38
5.1	Individual Customized GraphMAE2	48
5.2	Graph Representation of Fold-1	50
5.3	Graph Representation of Fold-2	51
5.4	Graph Representation of Fold-3	52
5.5	Graph Representation of Fold-4	53
5.6	Graph Representation of Fold-5	54
5.7	Confusion matrix for the ensemble model	56

List of Tables

3.1	Augmentation Parameters	22
-----	-----------------------------------	----

5.1	Hardware Specifications	45
5.2	Cross-validation performance metrics for the ensemble GNN model	54
5.3	Uncertainty-threshold performance metrics and coverage. . . .	56

Chapter 1

Introduction

According to the World Health Organisation (WHO), Cervical cancer, is the fourth most common cancer among women globally. Traditional therapies encompass surgery, chemotherapy, and radiation, as indicated by the patient's case study [18]. Over the years, cervical cancer has posed a serious threat to public health, more so in poor regions of the world where there is a shortage of trained personnel, especially colposcopists. Cervical cancer is preventable and is easily treatable for those women who seek early help. Nevertheless, it remains the most common cause of death from cancer in women. Cervical cancer alone takes 57,000 lives a year, with 30,000 deaths in a single year alone. In 2020, 604,000 new cases have been detected, and 342,000 deaths have taken place [23]. Screening methods such as visual examination of the cervix using a colposcope are very essential to identify precancerous lesions. Colposcopy is an early intervention strategy which can be considered to be a fruitful solution of early cancer detection. In colposcopy, a special magnifying device is used to look for abnormal areas of the cervix. The technique depends on the skill of the health worker. Also it is susceptible to human errors, and inter-observer variations lead different doctors making different conclusions regarding the same results. This variability can lead to differ diagnoses for the same condition, which can have implications for patient recovery. It may also present challenges in implementing proper screening practices in areas with limited resources.

In the past few years, artificial intelligence (AI), and more particularly deep learning, has been an effective force of change in the fields of medical imaging and also in disease diagnosis. Deep learning models like Convolutional Neural Networks (CNNs) have been exceedingly successful in analyzing complex visual patterns in various medical imaging modalities like histopathology, radiology, and dermatology. In cervical cancer detection, convolutional neural networks (CNNs) and transfer learning techniques have been very successful in classifying Pap smear and colposcopy images through hierarchical feature representations learning [54]. These techniques eliminate the dependence on human-engineered features and allow the models to generalize across different datasets. Moreover, pre-trained deep models such as ResNet, VGG, EfficientNet, and Inception have been successfully fine-tuned for cervical cancer classification with reduced training time and diagnostic accuracy being high [44]. Nonetheless, deep learning models are generally reliant on big annotated datasets, lack transparency in decision-making, and do not naturally process non-grid spatial relationships that are significant for medical images featuring complex and irregular structures [36].

In recent years, learning with Graph Neural Networks (GNNs) has drawn significant traction from various domains such as ranging from social network analysis and molecular chemistry to disease imaging [24]. GNNs are extremely proficient at learning complex relationships and dependencies in graph-structured data. As a result, they are an efficient tool for node classification, link prediction, and graph classification tasks. Motivated by the strengths of GNNs, this paper illustrates an efficient ensemble learning framework integrating different GNN architectures like GraphMAE2, GIN, GCN, and GAT models to improve predictive performance on colposcopic images. This ensemble method builds on the distinct advantages offered by each model. Consequently, it offers a very efficient method for dealing with complex and varied data categories. The major focus of this research is to further boost the diagnostic accuracy of stage-wise classification of cervical cancer by applying the ensemble GNN framework. We employ annotation to retain solely the superpixels that intersect with the cervix area, forming the Region of Interest (ROI), from each colposcopic image. The chosen superpixels are converted into nodes in a graph, which are then extracted to derive a collection of 75 features. In addition, these nodes and edges support complex and essential relationships. Our model’s interpretability was improved by adding an attention mechanism that weights each graph node to show how much it contributes to the final prediction. Our methodology includes the comprehensive preprocessing steps to ensure data quality and consistency, and then it is subsequently followed by strict training and evaluation protocols. We utilize Cross-validation and fine-tuning for ensuring maximum utilization of the ensemble model, with metrics tracking and visualization tools that provide us insights into the learning process. Furthermore, our framework incorporates uncertainty-estimating approaches that enable the model to assess confidence levels, hence optimizing the credibility of predictions. The experimental results ensure the efficiency of our ensemble framework, achieving a validation accuracy of 95.52% and an F1 score of 94.12%. By combining multiple graph architectural innovations, robust training and evaluation structures, and domain-specific enhancements (such as cervix-aware attention and uncertainty estimation), our research offers a strong framework at finding signs of cervical cancer from colposcopy images.

1.1 Research Motivation

Graph neural network (GNN) architectures, including Graph Convolutional Networks (GCN), Graph Isomorphism Networks (GIN), Graph Attention Networks (GAT), and more recently Graph Masked Autoencoders (GraphMAE2), have demonstrated exemplary potential in learning knowledge from complex, non-Euclidean data structures. The models are optimally positioned to identify complex relationships and interdependencies concealed in graph-structured data and thus optimally suitable for applications in biomedical imaging, where spatial and contextual information has paramount importance [55]. Within medical field applications, GNNs have exhibited promising outcomes across diverse tasks such as disease classification, prognosis prediction through effective modeling of interactions among cellular components. Recent research has initiated exploration of GNN applications in cervical cancer analysis, capitalizing on their capacity to integrate heterogeneous features and spatial relationships found in colposcopy images [50]. Nevertheless, the complete potential of advanced GNN architectures for this application remains insufficiently

investigated.

Our research was driven by the need to tackle the challenges associated with accurate and interpretable cervical cancer detection using colposcopy images. The Conventional image analysis techniques frequently fail to adequately capture the complex spatial relationships and subtle clinical characteristics present in the medical imagery. Drawing inspiration from the proven effectiveness of graph neural networks (GNNs) in processing relational data, we aimed to harness their capabilities for this particular application. Following these developments, we proposed a model that is the combination of the strengths of various GNN architectures in solving the complex task of detecting cervical cancer by using the colposcopy images. The method involved in graph creation from superpixels derived from colposcopic images. By developing a graph where nodes are the representative of salient tissue regions and edges are the representative of spatial interactions, the goal was to create a model that had the ability to distinguish local and global patterns applicable in disease detection. Additionally, we recognized the importance of clinical interpretability and reliability. This prompted us to include attention mechanisms to prioritize cervix-specific areas, node-level weighting, as well as to introduce uncertainty estimation, thereby enabling the model to convey its confidence in every prediction. By combining these advanced methodologies in an ensemble framework, our goal was to create a strong tool that not only yields high diagnostic accuracy but also delivers valuable insights and confidence scores to inform clinical decision-making.

1.2 Research Problem

Early detection of cervical cancer can improve the outcomes of patients' diagnostic imaging, particularly colposcopy plays a crucial part in this process. While traditional deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have had strong performance in the analysis of biomedical images [1], they do not capture the complex relational and spatial patterns of colposcopy images. Graph Neural Networks (GNNs), with the potential of modeling region-based relationships and of enhancing the feature extraction [16], offer an appealing solution. However, the use of graph attention mechanisms such as those of Graph Attention Networks (GAT) have not always resulted in better performance in the diagnosis of cervical cancer, according to current publications and peer reviews [6]. The key challenge is optimal use of attention mechanisms to selectively attend to clinically significant regions, especially the cervix region, while combining heterogeneous medical features. The main issue is that graph-based attention mechanisms are not being used to their full potential since they lack to do region-aware analysis, which is necessary to extract the most important morphological aspects. Current solutions sacrifice the importance of weighting the regions. Finally, there exists an urgent need for frameworks that can selectively attend to the cervix region and leverage self-supervised learning to enhance the representation of the features so the model can learn without the need for human labeling. In self-supervised learning, masked autoencoding aids in enhancing feature learning. Again, using the beneficial characteristics of several different GNN architectures by assembling can be a promising approach to achieve the optimal results in colposcopy imaging.

1.3 Research Objective

The aim of this study is to develop and evaluate a novel deep learning framework for the early detection of cervical cancer from colposcopy images, leveraging region-based, self-supervised graph neural network analysis. Firstly, we combine a variety of GNN architectures: GraphMAE2, GAT, GCN, and GIN into an ensemble learning scheme. Then we introduce cervix-aware attention scaling for improving the representation learning of both the nodes and edges of the built-up medical image graphs.

- **Region-Based Feature Integration**

Our initial objective is to focus mainly on nodes associated with the cervix region by leveraging a variety of medical features. This way, we can make sure that the most diagnostically significant points are highlighted in the graph representation.

- **Self-Supervised Graph Learning**

The second goal is to utilize a self-supervised graph neural network approach in the form of masked autoencoding (GraphMAE2) to enable the model to learn stable representations for nodes and edges through the reconstruction of masked features. Ensemble Graph Neural Networks involve creating an ensemble model that aggregates predictive power provided by various GNN architectures (GraphMAE2, GAT, GCN, GIN). This allows for comprehensive learning of graph-structured data and improves classification performance.

- **Cervix-Aware Attention Mechanisms**

Another objective of our research is to introduce and evaluate cervix-aware attention scaling as well as weighting node features proportional to their overlap with the cervix area, thus drawing the model’s attention to clinically significant regions.

1.4 Research Contribution

This contribution introduces a novel, clinically inspired graph-based method for cervical cancer risk prediction, distinguished by its integration of superpixel segmentation, region-of-interest (ROI) graph building, and advanced ensemble learning. Contrasting with earlier approaches that deal with images in their entirety or through patch-based CNNs [2], our approach segments each colposcopic image into superpixels, and these selected superpixels combine a biologically informative region. These superpixels serve as nodes of our graph in the region of interest, and edges are based on spatial proximity and tissue similarity, encoding local and global tissue architecture. Our derived graph-of-ROI framework provides the unique ability to model specific spatial relationships and interactions between heterotypic tissue regions, an essential element too often left out in previous work [10]. Each node is enhanced with a rich 75-dimensional colposcopy feature vector, which has been thoughtfully crafted to include acetowhite response, vascular pattern, surface texture, color, and lesion morphology, and thus provide a rich, informative representation of each ROI. We also leverage expert cervix zone annotations to inform graph construction and node weighting so that clinically significant regions receive

emphasis during learning. Our ensemble combines multiple state-of-the-art GNN architectures (GraphMAE2, GAT, GIN, GCN), each operating on these ROI graphs, to robustly capture multi-scale and multi-modal patterns. To our knowledge, this approach is first time done to unify superpixel-based graph construction, expert annotation, and high-dimensional medical feature extraction within an ensemble GNN framework for cervical cancer screening. This approach not only advances predictive accuracy but also enhances clinical interpretability, setting a new standard for explainable AI in digital pathology.

Chapter 2

Related Works And Background

2.1 Deep learning Model Used in Cervical Cancer Classification

Authors [26] introduce a deep learning-based approach to classify cervical transformation zones from colposcopy images, for which a number of researches had already been done in the literature. Determining the type of cervical transformation zone is important to grade the colposcopic diagnosis and also treatment policy. These newly developed sophisticated automatic diagnosis models include ColpoNet and CAIADS, which have remarkably high accuracies of 81.35% and 82.2%, respectively, regarding the monitoring of cervical lesions. Nevertheless, the detection of the transformation zone type is rather important for a proper diagnosis. It will first extract cervical detection of regions from colposcopic images through a Faster R-CNN model, able to find objects from images quite effectively. The proposed network, called GRF-Net for multiscale feature fusion, further processed such detected cervical regions. These extracted features from GRF-Net can then be used to classify the types of cervical transformation zones. Furthermore, it was observed that the proposed framework exceeded the state-of-the-art image classification models with an accuracy of 88.49% [26].

An article [48] shows an increasing interest in deep learning methods connected to medical image analysis for the screening of cervical cancer. A few related recent works in this respect have shown that deep learning algorithms, especially CNN, made great improvements in both performance and speed in the process of cervical cancer screening via colposcopic images. These further results showed that, though increasing the model classes was effective, there was an enhancement in an intra-dataset setting. The addition grossly reduced the performance on the test set. Accuracy during training, although high up to 99%, dropped down to 43.11% on test accuracy after increasing the number of classes. It also shows how different the specificity and sensitivity of various epochs are. Some classes in Experiments 1 and 2 had a few more images, and the results of both experiments truly did not make much difference in accuracy: 0.533 versus 0.524 after training with the same length-100 epochs. A Convolutional Neural Network(CNN) was proposed and validated by Hu et al. [5] to classify and detect anomalies on a large colposcopic image dataset. It can be attributed to much of the contribution that this model has had with regard

to high sensitivity and precision for cervical cancer screening. Other researchers applied CNNs for the categorization of colposcopic images as Normal and Abnormal, as discussed by Ghogh et al [7]. It finds its application in the diagnosis of cervical precancer, where packing multiple views enhances detection using multi-view CNN [8]. Sukegawa et al. [22] employed a multi-task learning framework along with CNN in classifying images as normal, precancerous, and cancer cases using shared representations. There is an advantage to its usage. Apart from that, methods that could realize high categorization precision in detecting cervical precancers included transfer learning. Li et al. [17] employed a pre-trained deep CNN from a natural image dataset and fine-tuned it with a cervical precancer dataset. This method has further been explored in the paper of Pramanik et al. [31], where the research has been compared with other DL techniques, including CNN and ensemble models.

Authors [11] review the recent developments related to deep learning technologies that have increasingly been applied to various diagnoses because of their self-learning nature. This goes hand in glove with the recent interest in a possible way that deep learning models further improve diagnosis accuracy by using medical images. The development of ColpoNet has taken place inspired by the DenseNet model owing to its computational efficiency compared to other models. This choice is representative of an emerging trend in the literature whereby deep learning models are created to be optimized for practical applications in healthcare. In this way, the performance of ColpoNet was benchmarked against several other deep learning models' performances, namely, AlexNet, VGG16, ResNet50, LeNet, and GoogLeNet, etc. For the initial processing, pictures are cleared, and this way it creates a clearer scan for the computer before the pictures can be analyzed by the ColpoNet system. This will lighten the images or, more likely, turn them slightly to have them viewed from the other angle. The end result of these will be that ColpoNet will become very adept at recognizing cancers, even if the images are slightly different every time. These experimental results [11] indicated that ColpoNet obtained an accuracy of 81.353%, hence outperforming state-of-the-art techniques. These were important steps in the research of automated cervical cancer screening. ColpoNet is mainly objective in trying to help with the early detection of cervical cancer cases, especially for places that do not have many ready specialists. It would make the whole process so easy that low-income countries would take center stage.

The authors [37] present a comprehensive review of the current techniques and technologies regarding the detection and grading of cervical cancer, particularly the challenge due to the presence of specular reflections in smart colposcopy images. In the conventional approaches of refilling the missing pixels of the digital images, there is a widely known state-of-the-art method called inpainting. Many techniques have been developed with a view to removing these kinds of specular reflections, but the process has scant automation. Most of the recent works using U-Net and GAN architectures of deep learning models that predict and remove glare in images come out promising in terms of their accuracy in identifying and segmenting the speculated reflection for quality improvement in cervical cancer detection. Authors [37] have highlighted how this forms a serious problem in the examination of cervical

cancer, whose grading is done into three CIN1, CIN2, and CIN3 classes. Grading will be carried out by the extent of abnormal cell growth, and only then is good quality imaging able to produce efficient treatment. Classifications are done using some state-of-the-art models like DenseNet121, VGG19, and EfficientNet. The performance of all these methods is tested on enhanced and unenhanced images. The achieved accuracy is 97.32% for DenseNet121, VGG19 of 96.25%, and finally, Efficient Net, which stands at 96.75%. These assure that the approach being used here improves the grading operation concerning glare issues.

An article [42] upholds a wide discussion of critical issues related to diagnosis, particularly on cervical cancer and generally speaking on the integration of deep learning techniques. It expresses the fact that, though this ailment is almost preventable, yet the mere presence of current screening processes cannot help achieve that target. This is because it gives an elaboration on the various methodologies applied in diagnosing cervical cancer using deep learning techniques, mostly around cytology and colposcopy images. Different deep learning models used in diagnosing cervical cancer are elaborated on. Among them, the most used models in the work of [42] are VGG-16, ResNet (50-101), EfficientNet (B0 to B3), and Inception V3. The models make a promising set with good accuracy rates. Whereas in cytology models, it was around 96%, for colposcopy, the average stood at 91%. Most recently, there was the development of another neural network classifier in colposcopy, that included the use of a hybrid algorithm. Feature extraction uses EfficientNet-B0, assisted in efficiency during training by a Gated Recurrent Unit. This kind of methodology has so far been presenting better performance than conventional neural network architecture. Authors [42] point out that there are limited numbers of openly available cervical screening datasets. With few images, this presents a great challenge to train and validate an AI model. Therefore, poor model performance under real-world application conditions arises. Authors [42] discusses obstacles in the execution of AI into cervical cancer detection. Therefore, more research as well as development is still required.

Author [15] reviews the contribution to the design of a new multi-modal diagnostic system capable of classifying, segmenting, and detecting classifications on colposcopy image analysis. This system supports the doctors for diagnosis and biopsy to consider high-grade squamous intraepithelial lesions. In view of the above, for the proposed methods with the use of ResNet, two-class classification of the colposcopy images is done, with the exact segmentation of the lesions by the U-Net model for classification and division. Therefore, regarding the segmentation model, 11,198 acetic and iodine images were used for training a huge dataset to correctly identify biopsy sites. The proposed classification model in the paper reached an accuracy of 84.10% for ordinary images. Another big outcome of this model was the very high AUC of 0.9261, thus qualifying for primary triage in colposcopic examination [15]. It has also, through most of its features related to the brightness and saturation of the image, been able to show that this variation in diagnostic accuracy was 84.10% for normal and 63.83% for high-definition images. Author [15] thinks previous studies are lacking in that they are usually of a small sample size and geared toward

simple classification. This covers the significant gap in gynecological diagnosis and extends the application of deep learning.

The main obstacle to the development of AI-based diagnostic systems is the unavailability of properly annotated medical datasets. Most of the works used publicly available datasets like Kaggle and IARC, which contain less than 5000 labeled images. The dataset proposed in paper [35] consists of a total of 456 colposcopy images gathered from the VIA Image bank and Colposcopy Image Bank. The best Naive Bayes Classifier reached 69.34%, while the Simple Logistic reached an accuracy of 76.64% using GoogLeNet. In this way, among these, the random forest has done very well with an accuracy of 71.53% using ResNet-18 with transformation zone extraction. One can see from here the rate of accuracy is too low. It can be because of the small dataset. Another article [34] has proposed a small dataset consisting of a total of 448 images, of which 128 are colonoscopy images and 320 are patches of Pap test images. The methodology section explains how the ensemble approach has been impacted in the merge in DeepColpo for the detection of colposcopy and also DeepCyto+ for cytology images. KNN did classify the images into normal and abnormal well, indicating that the ensemble of several models may result in very good accuracy in diagnosis. DeepColpo, once fine-tuned, gave results against a test with an accuracy of 63.64%. Simultaneously, that means the small dataset is one of the major obstacles toward gaining more accuracy that could generalize to more problems.

The Study [43] mainly focuses on enabling initial detection of cervical cancer and hence proposes a method for the categorization of cervix images using deep learning and machine learning techniques. It functions on the principle of the ML classifiers on the pre-trained models like ResNet101, ResNet152, and AlexNet for giving the best diagnosis of the cervical cancer, that was possessed by the prevailing detection algorithms. In the paper [43], there were 4068 pap smear images. From this data, 80% was used for training while 20% was utilized for testing. It was also among the main features that assured the reproducibility of the study, showing performance of different models against it. Data from DL models were used for training Machine Learning methods-namely, the Simple Logistic, Random Forest, Naive Bayes, Decision Tree, and Principal Component Analysis-which sorted the data into groups. This hybrid model retains deep learning flexibility for feature removal and retains machine learning for the associated categorization tasks. The best performance, by Simple Logistic over ResNet152, resulted in an optimal accuracy of 98.08%. Other models, such as ResNet 101 or even AlexNet, were doing very well in this regard, but lagged behind as compared to ResNet152 because it showed higher accuracy. The result may not apply in more clinical settings since it is small, comprising only 4068 images. When this size of data increases, things are bound to get even better and make model reviews reliable. The work [43] outlines that for any further improvement in the precision and stability of cervical cancer categorization, there is a need to use larger and more heterogeneous datasets and also employ more advanced techniques for ensemble learning.

An article [29] is on the topic of using deep learning methods to diagnose some cervical conditions like chronic cervicitis and cervical cancer. According to the paper, screening precancerous lesions is unavoidable to prevent the appearance of lesions resulting in cervical cancer. It goes on to say that various deep-learning-based technologies, especially CNN, have been forging ahead with rapid growth. Having said so, this technology has been found quite assuring in automating diagnoses for a range of medical conditions, including cervical cancer. This technology will continue to enhance the diagnostic accuracy and reliability of colposcopic image analysis. In this paper, the dataset used was from Hunan Provincial People’s Hospital, including 10,012 colposcopy images taken from 1,081 patients. Preliminary classification of such images, due to high similarities of some conditions, returned low accuracy of 42.16%. Since it has been regarded as containing two major features: chronic cervicitis and cervical cancer, the overall impressive performance was 95.19%, proving the potential of deep learning in improving diagnostic results. These facts would, therefore, give substance to the fact that the provided deep learning model is efficient and promising for applications in diagnosing cervical conditions for the benefit of the patients and in order to facilitate the procedure of the screening more easier.

The article [40] focuses on a low-resource alternative with the DL model, with improved precision in cervical disease diagnosis. The important contribution of this paper [40] is to present a prediction model using CNN algorithms. In all, an Intel & MobileODT Cervical Cancer Screening dataset of several types of cervical images was used for training and testing of the CNN model. In this paper [40], the authors have taken 6,734 images labeled from the Intel & MobileODT Cervical Cancer Screening dataset. Furthermore, this dataset was divided into three categories involving three CIN stages: CIN1, CIN2, and CIN3. Authors proposed a model through which features of cervical images could be extracted effectively by making use of the VGG-16 architecture that comprises sixteen layers. It is above 92%, where the best is an experiment with 99% accuracy during the training. This therefore indicates that it’s very capable of learning from the training data and generalizing very well into unseen data. Thus, this addresses some of the limitations characterizing most current diagnostic methods, since they cannot satisfactorily detect a number of cervical diseases. This paper tries to overcome some limitations of the previous works by utilizing deep learning for improved diagnostic aid to healthcare professionals.

To increase the accuracy of the diagnosis models, ROI extraction plays a significant role. The transformation zone forms a part of the Region of Interest and it forms the key area where most of the lesions take place; hence, this section is quite crucial and important for any diagnosis against the background of rich information on the detection of cervical cancer. Preprocessing methodologies, which enhance the contrast of the original images, allow better visibility of the transformation zone. Such high performance of the given diagnostic models mentioned in paper [35] is based on utilizing already pre-extracted transformation zones in the images-for example, the most accurate one, 76.64%, was performed by a simple logistic classifier with the feature extractor of the GoogleNet model. This proves the high accuracy of the

proposed method of in-focus area extraction in classification. Various challenges in dealing with entire colposcopic images full of meaningless information that may not be needed in the analysis can be read from another paper [35]. It would not serve any purpose to handle all of these images directly, which may be computationally complex and degrade the performance of the model drastically. Thus, the exact ROI extraction does not stimulate an advantage but rather improvement of the diagnostic model. Besides, paper [27] carried out the annotation of the cervix image in a simple way with the help of the Dark Label annotation tool. For the extraction of ROI, a MobileNetv2-YOLOv3 model was used, for extraction-a 99.89% mAP was obtained, testifying to the very accurate extraction of the region of the cervix. In their original dimensions of the paper dataset images, they were 416×416 pixels; hence, in the process of down-sampling, dimensions 13×13 , 26×26 , and 52×52 are achieved after which the output images are allowed to train detection networks. Further, the downsampled network coupled with other pre-existing networks, which classify the transformation zones. Here, two dense feature fusion network layers are used, where one dense block is utilized in the first layer and 2 dense blocks are utilized in the second layer. In this paper, DenseNet uses strides of 32, 16 and 8 to get the output image dimensions 13×13 , 26×26 , and 52×52 , respectively. For semantic segmentation networks, the Adam optimizer was implemented with an initial learning rate of 0.001; the optimization and training used a batch size of 8. Moreover, the EfficientNetB0 model, besides the other deep learning models using ROI-extracted data, is also cropped and gave a best result in the paper on validation accuracy of 92.19% [26]. Contrasting that, paper [26] reaches an accuracy of 88.49% in the classification test; therefore, the multiscale feature fusion test identified better areas in the cervix for greater precision.

An article [45] introduces WFC2DS (Web Framework for Cervical Cancer Detection System), a new web framework for early diagnostics of cervical cancer, a major women's medical issue in general worldwide. With the help of advanced machine learning classification algorithms, this framework processes information of patients for increased accuracy and medical care in diagnostics. The research focuses on a sequential methodology, starting with a thorough analysis of a 858-patient record, 36-attribute collection, then reduced to 31 useful features [1]. Various algorithms such as K-Nearest Neighbor (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM), and Decision Tree (DT), have been utilized for information classification. Venezuela's Universitario Hospital de Caracas' 858-patient record, 36-attribute collection involves demographics, behavior, and medical information. As a moderately big collection, larger collections could maximize model performance and generalizability. In this study, High accuracy values have been attained, with accuracy at 98.1% (F1: 0.98) for RFC, closely in performance with Decision Tree, AdaBoost (97.4%, F1: 0.98), ANN (97.7%, F1: 0.96), SVM (96.2%, F1: 0.96), and KNN (90.6%, F1: 0.91). While in this paper, no specific discussion of limitation mentioned, problems could involve restrictions in terms of dimensions in a collection, constraining model training and model testing, and generalizability concerns due to all data are based on a single sourced information derived from a single institution. These problems come from the fact that it's hard to get enough data, and the need for bigger datasets to make models more reliable and prevent overfitting.

To sum up, WFC2DS model approaches cervical cancer diagnostics with significant improvements in its accuracy and medical care in diagnostics, supporting early cervical cancer diagnostics for fewer deaths and improvements in female medical care.

Authors [14] highlight the major challenges facing cervical cancer screening. The limitation in the conventional colposcopy as outlined in LMICs, which includes properties such as operator-dependent, variation in diagnostic performance, and lack of knowledgeable clinicians producing problems in suboptimal detection of CIN and a related increase of risk of poor patient outcomes. In fact, WHO has targeted the elimination of cervical cancer by means of increased screening and appropriate management by 2030. However, LMICs are plagued with poor infrastructure and a shortage of trained personnel. Overall diagnostic accuracy in LMICs for colposcopy varies from 30% to 70%. Poor operator experience and shortages of service capability are the major reasons for low diagnostic accuracy. AI-driven digital colposcopy improves diagnostic ability by automating image analysis and reduces subjectivity in colposcopic diagnosis. This is because it integrates AI algorithms in digital colposcopes, hence enabling real-time reporting, improving the clinical workflow in resource-poor settings. AI might also act as an educational tool, training the novice colposcopists and filling the competency gap in LMICs. Important works have been done to develop AI algorithms to analyze cervical images, promising improvement along many aspects: efficiency and diagnostic accuracy. Poor management of colposcopic image datasets in general ensures poor labeling and quality control, a major barrier to both training and validation of AI models. Generally, the interpretability of AI models remains another challenge. If the colposcopists are to trust and use the diagnostic tool, then they have to get transparent explanations of the AI diagnostic decisions. A cloud-based AI platform would democratize access to diagnostic tools throughout LMICs, hence overcoming health disparities. Collaboration among research institutions is crucial for translating AI algorithms into clinical practice, and supporting global efforts to eliminate cervical cancer.

2.2 Graph Architecture Utilized in Medical Image Diagnosis

An article [41] critically writes about graph-based cervical cancer segmentation methods and evaluates how they have developed and where they fall short with respect to efficient diagnostic and therapeutic planning. The article summarizes evidence from 20 studies based on various databases, such as the Herlev database and cervical biopsies without mentioning dataset sizes as a mixture of large and small datasets. The methods addressed in the discussion include specialized graph-based techniques, namely superpixel-based techniques, with enhanced segmentation accuracy achieved by combining structural details such as colour and texture. The article fails to specify accuracy levels for the methods but hints at the need for large-scale clinic validation to evaluate performance at a real-world level. Obvious shortcomings in the article include failure to exercise comparative performance evaluation as

well as intensive use of high-resolution imaging and challenges in segmenting tissues with high accuracy. The reasons provided by the article were the intrinsic heterogeneity of cervical tissues and the limited volume of labeled data. The article [41] concludes by urging future studies to adopt state-of-the-art deep learning techniques perform high-quality clinical validations, and improve the interpretability of results in order to overcome these core challenges in generating reliable segmentation tools for cervical cancer care.

The authors [28] introduce a new self-supervised learning method for graph data with a target of generative SSL with masked graph autoencoders (GAEs) in order to overcome the weaknesses of current methods, weak objectives of reconstruction, and training robustness. The new method adopts a masking procedure in which a portion of node features is masked and reconstructed by the model as a form of a scaled cosine error for enhancing training stability. The model was tested with extensive experimentation on 21 public datasets, Cora, Citeseer and PubMed, which were on a vast scale in terms of possessing diverse graph structures and properties. GraphMAE was competitive and outperformed state-of-the-art baselines in a variety of tasks comprehensively and demonstrated being strong and efficient in terms of node classification and classification of graphs. The authors [28] also elucidates the weaknesses of the study, and potential trivialized representations because of overly simplistic designs in state-of-the-art GAEs, which may affect performance on the basis of dataset properties. The weaknesses are a consequence of the intrinsically demanding task of representation of a graph and the task of successful extraction of features. Overall, GraphMAE is a landmark in self-supervised learning of graphs and has promising potential for usage in this domain.

An article [33] introduces a new architecture capable of addressing the label scarcity problem of graph data with the help of masked graph autoencoders. This article [33] discusses two broad approaches: multi-view random re-mask decoding to introduce randomness into the process of reconstruction and prediction of latent representation to impose embedding-space reconstruction and influence the quality of feature reconstruction. The method includes comprehensive experiments on large public datasets like ogbn-Papers100M with 111 million nodes and 1.6 billion edges and brings into focus the performance of the proposed method in a large scale. GraphMAE2 achieved state-of-the-art performance with considerable improvements, with a 2.45% improvement in performance compared with state-of-the-art methods in the OGBN-Papers100M dataset. The work, however, has a few weaknesses, like the fact that GraphMAE2 and GraphMAE outperformed a randomly initialized baseline in the ogbn-Papers100M dataset but failed to outperform the other contrastive baselines as expected. These weaknesses uphold a significant way regarding the significance of appropriate data augmentation techniques in self-supervised learning. GraphMAE2 constitutes a great innovation in the domain of graph self-supervised learning, providing important information and being a source of inspiration for future work.

An article [50] deals with the classification of cervical lesions, which is essential for the early detection of cervical cancer. The authors introduce a new framework that integrates colposcopy images, segmentation masks, and graph representations for improved classification outcomes. The authors [50] formulated a Graph Neural Network (GNN) architecture and assessed it with a five-fold cross-validation. The model, upon fine-tuning, achieved a macro-average F1-score of 94.56% and also a validation accuracy of 98.98%. The data in this study included various images from different classes, i.e., normal and carcinoma, a total of 1,196 images. Image ground truth masks were also created by the authors through a custom script in Python, which read JSON annotation files for accurate segmentation. Despite the favorable results, the study identifies a number of limitations. The segmentation model is sensitive to the quality of masks generated by DeepLabV3, which can introduce errors if not refined from misclassified pixels. Graph building is computationally expensive, especially with big databases. The model must adapt numerous hyperparameters for generalization on different colposcopic databases. Overall, the proposed framework is promising in clinical utility but is not without some potential for optimization. Future studies can explore self-supervised learning approaches in order to achieve performance with fewer manually annotated pixels.

The article [38] is a comprehensive report on the historical development of cervix cancer detection techniques. From traditional methods like the Pap smear test to cutting-edge computer-aided detection (CAD) systems, in the wake of the need for higher accuracy, sensitivity, and specificity for screening. The study [38] undertook a systematic review approach, carrying out a search in the Scopus database for works that had been published between 1996 and 2022, and had included 108 works in the review with certain standards. The article does not present a clear picture of the actual dataset size that had been included in such studies, and remains a pending issue. The review highlights that even as early CAD systems suffered from low sensitivity and specificity, recent innovation with the help of machine learning (ML) algorithms had achieved accuracy levels as much as that of up to 98.3%. In spite of such progress, the article is also limited by the fact that CAD systems' effectiveness and credibility continue to be tested owing to the unsubstantiated nature of traditional methods as well as probable false negatives in early screening. Overall, the article highlights the long progress in cervix cancer detection techniques and underscores the necessity for ongoing innovation and research with the aim to further develop such techniques and help bring relief from the burden of the disease for women who suffer from it all over the world.

An article [13] suggests a new method called Improved Normalized Graph Cut with Generalized Data for Enhanced Segmentation (INGC-GDES) to help detect cervical cancer from pap smear cell images. It is designed to enhance the detection of the cytoplasm and nucleus boundaries, which is important for the early detection of cancer. As delayed detection can decrease the survival rate in women. The researchers [13] use a generalized eigenvalue-based method to improve the segmentation of images. It also uses a maximum flow algorithm to estimate the best pixel points for the detection of cancerous cells. The study uses the Herlev dataset, which

consists of varied images of cervical pap smears. But it does not mention the size of the dataset, so it is not clear if it is small or large. The suggested method shows a significant improvement in classification accuracy with a 28% improvement over existing methods. It also enhances sensitivity and specificity. The paper does not clearly mention the limitations, but the possible limitations can be dataset size and variety, which can affect the outcome. The limitations can be due to difficulty in obtaining a large enough dataset for training and testing. Overall, the INGC-GDES method is effective in reducing noise and enhancing boundary detection and has the potential for better detection of cervical cancer. Future studies can be aimed at optimizing the methodology even more and enhancing any limitations in the study.

Authors [21] suggest a new cervical cell classification method using a Graph Convolutional Network (GCN) for improving the accuracy of Pap smear image-based detection of abnormal cells. The main goal here is the improvement of cervical cancer screening through capturing the underlying relationships between cells of the same category that traditional CNN-based methods do not capture. The process begins with image feature extraction using a CNN model, followed by clustering of similar cells and graph structure construction from the clusters. The inter-cluster relationships are learnt by the GCN, enhancing the classification accuracy. The two datasets employed in the work comprise the popular SIPaKMeD public corpus of 4,049 images containing 5 classes, and a new one comprising a larger-scale Motic liquid-based cytology corpus spanning 7 classes. This works out to a relatively large corpus size overall. The proposed approach outperforms previous methods with regards to accuracy, sensitivity, and performance metrics in general, although the actual percentage accuracy figure isn't given. One weak aspect of the study would be the possible influence of staining variation and clustering efficacy, and consequently, the impact on the metrics. These would naturally arise from variation in image and label quality across the datasets. Overall, the method shows tremendous potential for the improvement of automatic cervical cell classification and could potentially find applications in actual cancer-screen-based systems.

The article [32] introduces a new method for detecting cervical cancer with an Improved Boykov's Graph Cut-based Conditional Random Fields and Superpixel-imposed Semantic Segmentation Technique. The method seeks to improve the process of segmenting cells that, in Pap smear tests, play a vital part in identifying cancer in the early stages. The method incorporates several approaches to improve the definition of the boundaries of the cell structure so that the process of demarcating the normal cells from the cells that are not normal (abnormal cells) becomes easy. The article [32] fails to state the actual volume of the data that had been employed, but requires a high-strength dataset to adequately train the model. The proposed method had an accuracy rate of 99.78% with a consumption of only 2.18 seconds and had high precision, sensitivity, and specificity levels as well. The authors therefore do suggest that the methods being used currently come with the staining inconsistency on Pap smear tests that would influence the segmentation process to become less accurate. The limitations are actually brought about by the inconsistency of the image and staining processes employed in the clinical setup.

The article [32] concludes that the method has high potential for detecting cervical cancer but that the staining issue needs to be resolved for the method to achieve effectiveness in real-world situations.

Author [20] suggests a new framework for cervical cell classification by leveraging Graph Convolutional Networks (GCN). It aims at enhancing classification efficiency by exploring relations among images. An advancement from traditional Convolutional Neural Networks (CNN), based mainly on deep features. There is an application for this research on early cervical cancer screening. There are two datasets on which the researchers base their study: SIPaKMeD dataset with 4049 single-cell images and Motic dataset with many images and 7 classes. There are plenty of images within these datasets and are beneficial for training and testing the proposed framework. An enhancement of classification performance is evident from the GCN framework. According to this article [20], although no rates of accuracy are presented. There are limitations to this study, however these are reliance on quality of expert labels and vulnerability to overfitting due to complexity of model. These can result from variability of expert labeling of data and need for a large dataset for efficiency of model. However this article clearly suggests a framework with much potential for enhancing cervical cell classification and contributes notably to automated screening methods and cervical cytology.

One of the principal limitations of traditional supervised methods is that they require large, labeled datasets, which are costly and time-consuming to obtain [30]. The article [12] offers a GCN-based semi-supervised learning method that dynamically modifies the adjacency matrix during training. It outperforms traditional supervised methods like SVM and ResNet-18, especially in low-label scenarios. The authors [12] discovered the proposed method capable of significantly enhancing the classification accuracy for cervical intraepithelial neoplasia on labeled and unlabeled cervigram data. The paper describes the limitations of existing models, particularly Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT). GCN treats all the neighboring nodes equally, potentially ignoring strong correlations [4]. GAT, on the other hand, only takes 1-hop neighbors into consideration, missing useful information from distant nodes [47]. The authors [19] introduce the Gated Graph Attention Network (GGAT) that involves a gating mechanism in combination with attention mechanisms to enhance the model’s capacity to learn multi-hop neighborhood information. The innovation aims to improve the accuracy of cancer prediction by a significant margin.

An article [25] discusses various pooling methods, highlighting inefficiencies of traditional approaches like global mean or max pooling in handling class imbalances of WSI (whole-slide image) data. Most of the previous work lags behind the potential of providing attention mechanisms in graph models, resorting instead to more computationally costly transformers [49]. This shortcoming opens up incentive for the exploration of recent spatial graph models with attention towards WSI analysis. Authors [52] introduce the Projection-gated Graph Attention Network (Pg-GAT) as

a solution to these shortcomings by utilizing graph structures with attention and employing differentiable projection-gated pooling for rendering the model efficient and interpretable. [52] include failure instances where the model did not perform, specifically highlighting that 12-hop depth was insufficient in very large tumor regions. This shows the need for further tuning of hyperparameters for better performance in such instances.

Chapter 3

Methodology

3.1 Structure of Our Proposed Model

Firstly, we collected 1333 colposcopy images from our integrated dataset. 1279 images out of these were retained later as many fields had missing values so we labeled the dataset on HPV status for binary classification. Our images were resized to 800*600 as a standardization step. Then, we divided each image into superpixels using the SLIC method. Then, we extracted 75D colposcopic features from every images. Again, we used the makesense.ai website to annotate the cervix zone first and then labeled it with the HPV status, collected from the metadata. This tells us how much the superpixels overlaps with the cervix region, helping us build the ROI(Region of Interest) from every images. This ensures the region's features are weighted. The weight(attention value) is higher on the cervical region than the rest of the picture. These superpixels turn into the nodes of the graph and edges between the nodes evaluate their relationships and feature values. Moreover, all the images were splitted into train(60%),test(15%),validation(25%) subsets. In addition, we used 4 different models(GraphMAE2, GIN, GCN, GAT), customized and integrated them together to send into supervised training stage. But in the pretraining stage, we sent the GraphMAE2 model with masked autoencoding. The pretraining stage runs for 30 epochs. The graph data is then passed through the encoder and classifier and output is compared to the true labels using a loss function. Here, the data is split into multiple folds. For each fold, a separate model is trained and validated. The best models from each fold are saved. After each epoch, the model is evaluated on the validation set to monitor performance (accuracy, F1, etc.). If validation performance does not improve for a set number of epochs (20 epochs), training stops early to prevent overfitting. In every epoch, the model with the best validation F1 score is saved if fine tuning does not give better results. All the metrics are calculated along with the uncertainty and prediction. Our proposed architecture workflow is given below for understanding.

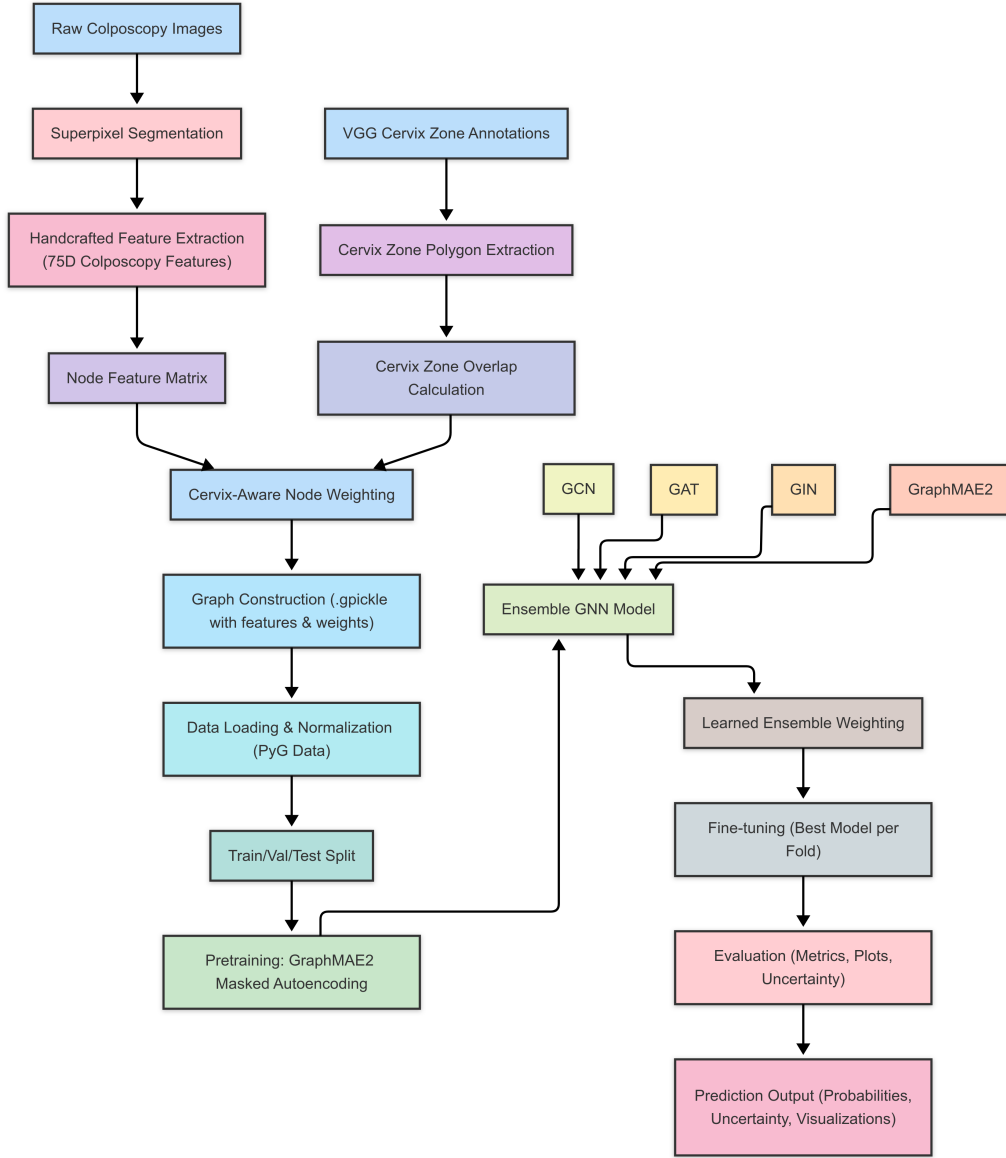


Figure 3.1: Proposed Model Workflow Diagram

3.2 Dataset Collection

We conducted our research using a colposcopy picture based two datasets with an emphasis on cervical cancer detection and categorisation. We used these datasets as these proves itself to be a consistent dataset that will improve the methodologies used for automated screening for cervical cancer. Our first dataset is IARC Visual Inspection with Acetic Acid (VIA) Image Bank which is obtained from WHO(World Health Organization) . It includes 180 cases and 420 images. The other dataset, IARC Colposcopy Image Bank which was also obtained from the WHO (World Health Organization). It includes 200 cases and 913 images. After combining the two datasets we got a total number of 1333 images in our combined dataset.

The colposcopy images in the dataset ensure adherence to recognized colposcopy protocols. They also guarantee uniform quality across all samples. It also included clinical documentation of patients of cervical cancer. This metadata was routinely

collected from the first dataset, IARC Visual Inspection with Acetic Acid (VIA) Image Bank, and recorded in "Cases Metadata.xlsx". It is an expansive collection that also has expert clinical observations along with patient details. It is included with our colposcopy image dataset to aid in knowledge development. It is mainly the most important component for our research for constructing a context-aware cervical cancer detection system. The substantial clinical metadata of the dataset includes 16 essential features for every single case. These characteristics are columnar epithelium, following adequacy, metaplastic squamous epithelium, squamocolumnar junction visibility, transformation zone characteristics, original squamous epithelium, location of the lesion, Grade 1 and Grade 2 classifications, suspicious invasion indicators, aceto uptake patterns, margin characteristics, vessel patterns, lesion size measurements, iodine uptake responses, and SwedeFinal scores. This rich metadata enables us to have multi-factorial analysis for our exploring in the research.

3.3 Dataset Pre-Processing

3.3.1 Dataset Standardization

There was a metadata file with clinical and diagnostic information for each case that came with the first dataset. The original metadata was made of 43 columns that held a wide range of clinical and diagnostic information. But not all columns had full data for every image, even many fields had a lot of empty or missing values, which meant they couldn't be used for analysis. Also, there were very few pictures for some diagnostic categories, sometimes just one or two, which caused major problems with class imbalance. We excluded any category with fewer than 10 images to make sure that the distribution of the classes was fair and to keep the statistical validity. A total of 1,279 high-quality pictures were retained after this filtering process. We used these filtered images and their HPV status to make a new, clean metadata file that we could use for our research. We used the HPV status column for labeling because it gave us consistent and full information about how to classify all of the samples we kept. At first, the pictures had two sizes: 800x600 pixels and 920x720 pixels. Since the higher resolution (920x720) didn't show up consistently, all the images were resized to 800x600 pixels as a standardization step. This made sure that all the input measurements were the same.

3.3.2 Dataset Splitting

The combined dataset of 1279 images has been divided into three subsets: training, testing, and validation. Because our dataset is small, we decided to use 60% of the training set and 25% of the validation set. When we utilized 15% for validation and 70% for training, the validation set had insufficient sample size. As a result, we were having overfitting concerns. We wanted to optimize the balance between the train and validation sets. We allocated 15% for testing, 25% for validation, and 60% for training. The training set comprises 776 images, representing 60% of the overall dataset. The test set has 187 images, which is 15% of the total dataset images. Again, the validation set includes 25% images of the whole dataset, which is 316 images in total number. From the metadata that was provided with the initial dataset, we extracted the image IDs along with their corresponding HPV

classification labels. This dataset is mainly focused on binary classification of HPV (Human Papilloma-virus) infection, with each image labeled as either HPV-positive or HPV-negative. In the matter of class distribution, the training set includes 386 HPV- positive and 390 HPV-negative images, while the validation set consists of 151 HPV- positive and 165 HPV-negative images in total. The test set contains 89 HPV-positive and 98 HPV-negative images. Additionally, the dataset comprises 363 cases in total, with 217 cases allocated to the training set, 91 to the validation set, and 55 to the test set. This distribution guaranteed a balanced representation of both classes across all subsets of images while maintaining the integrity of individual cases.

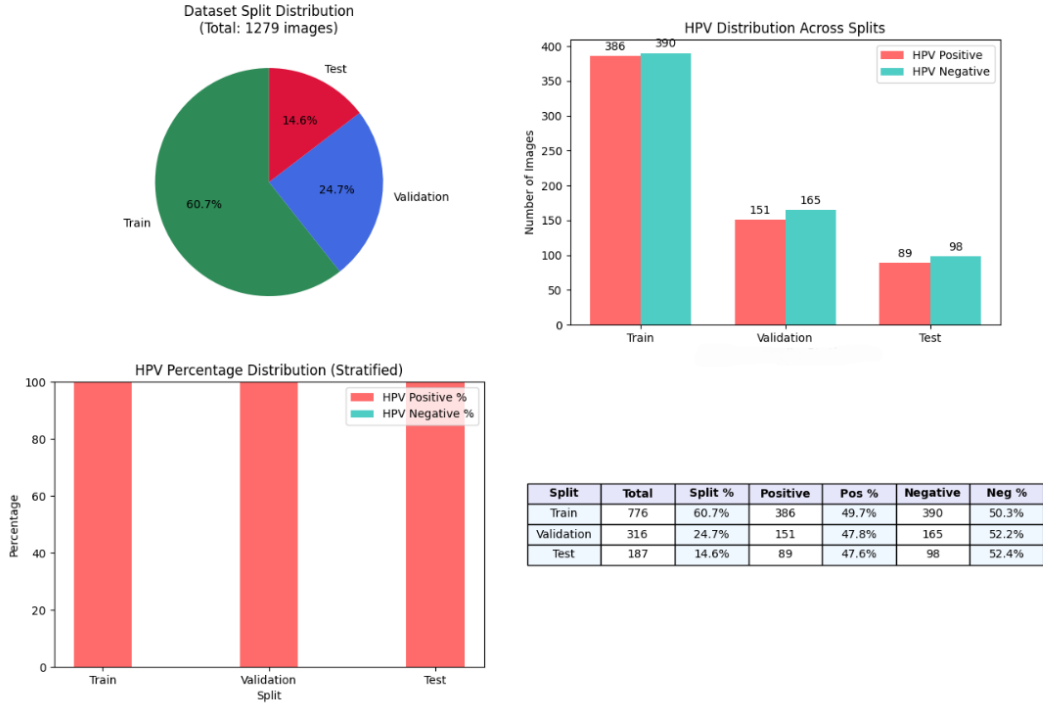


Figure 3.2: Dataset Split and HPV Distribution

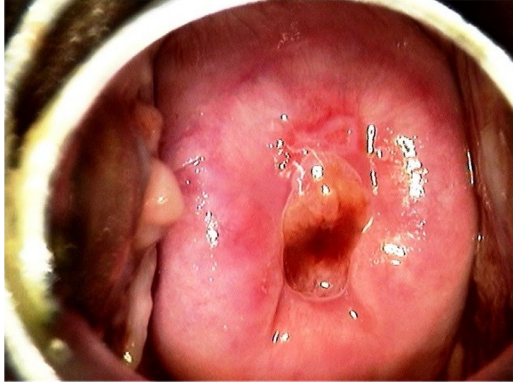
3.3.3 Dataset Augmentation

To prevent the problem of overfitting and improve the model’s generalization ability, data augmentation was implemented solely on the training subset. Each image underwent two rounds of augmentation employing careful, clinically suitable techniques to maintain diagnostic integrity of the dataset. The augmentation configuration was carefully planned to prevent any alterations that might compromise structural features. A limited rotation range of ± 15 degrees along with horizontal flipping was implemented to mimic natural variations in camera angle and viewpoint. A minimal amount of Gaussian noise (0.1) was introduced, considering that the original images already contained some noise. The augmentation emphasized subtle lighting and contrast modifications to replicate the variability encountered in real-world clinical settings. These alterations included several things. For example, Controlled brightness shifts within a $\pm 5\%$ range, contrast adjustments between 0.9 and 1.3, and gamma corrections ranging from 0.8 to 1.2 were all part of the adjustments.

We carefully adjusted it to match changes in natural lighting and reduce reflection sensitivity. CLAHE (Contrast Limited Adaptive Histogram Equalization) was implemented with a clip limit ranging from 1.0 to 3.0 for the purpose of enhancing local contrast. A subtle sharpening effect (strength: 0.1–0.4) was implemented to mitigate blur, especially in the scenarios involving low-resolution images. The variation in saturation was minimal, spanning between 0.95 to 1.05. Again, hue shifts were confined to a narrow $\pm 3^\circ$ -degree spectrum, thus retaining diagnostic color fidelity. All augmentations were implemented with a probability of 0.5, ensuring a balance between randomness and the preservation of the original data distribution’s credibility.

Parameter	Value/Range
Rotation Range	$[-15^\circ, +15^\circ]$
Horizontal Flip	True
Noise Level	0.1
Brightness Range	[0.95, 1.05]
Contrast Range	[0.9, 1.3]
Gamma Correction	[0.8, 1.2]
Apply CLAHE	True
CLAHE Clip Limit	[1.0, 3.0]
Apply Sharpening	True
Sharpening Strength	[0.1, 0.4]
Saturation Range	[0.95, 1.05]
Hue Shift Range	$[-3^\circ, +3^\circ]$
Augmentation Probability	0.5

Table 3.1: Augmentation Parameters



Augmented 1



Augmented 2



Original

Figure 3.3: Moderate Augmentation

3.3.4 Annotation

Getting information from the cervix zone was a crucial part, so we used the make-sense.ai website to mark the cervix zone ourselves from the images using a polygon, and also labeled them according to their HPV status from the metadata. After annotating each set (train, test, validation), we downloaded the information using the VGG format.

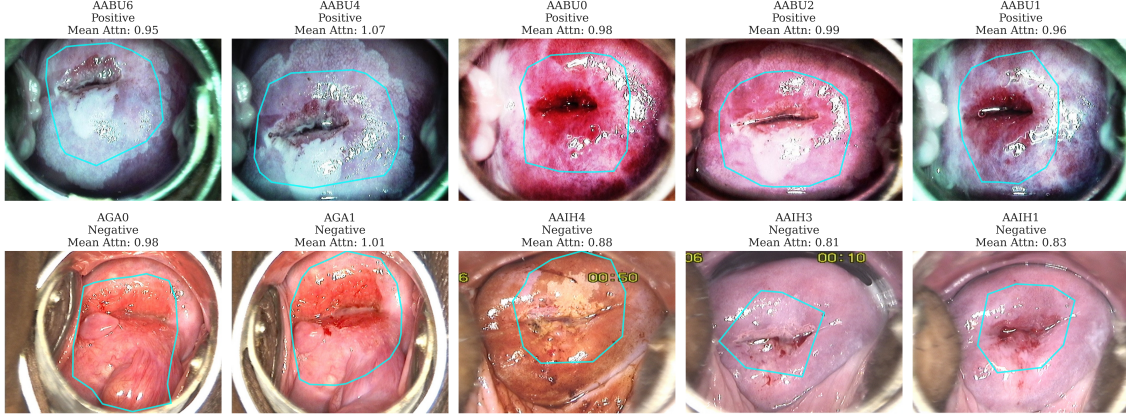


Figure 3.4: Cervix zone annotation from colposcopy image

3.3.5 Feature extraction

To support HPV detection and better understand the surface of the cervix from colposcopy images, we developed a specialized method that extracts 75 different features from each image. These features capture important visual details that doctors typically examine during a colposcopy and are grouped into five main categories. Initially, we process the images by transforming them into various color formats such as grayscale, HSV (hue, saturation, brightness), and LAB. This enables us to identify a broader spectrum of visual patterns. The initial set of features centers on acetowhite lesions, which shows up as white regions following the application of acetic acid to the cervix. These features measure how bright and widespread these white areas are, how many such spots exist, what shapes they take, and how sharply their edges appear. A total of 20 features are collected from this step. Next, we analyze the blood vessel patterns, which are important for spotting signs of abnormal tissue. We seek to identify patterns such as small dots (punctuation), tile-like areas (mosaicism), as well as irregular branching. Utilizing advanced image filters allows for the enhancement of blood vessels, enabling the measurement of their density, shape, and orientation. This contributes to an additional 20 features. In the third step, we measure the texture of the cervix surface. As rough textures can sometimes suggest problems, smooth textures often look normal. To analyze this closely, we use methods that detect how patterns repeat, how sharp or soft the surface looks, and how directional the texture is. From this part, we get 15 distinct features. Subsequently, we examine the attributes of color and tissue. We measure average color values, brightness, how much colors change, and even the color temperature (the balance of warm and cool colors). This is because different tissues reflect light in different ways. In this way, we can tell the difference between healthy and unhealthy tissue by these 12 features and identify them. Finally, at the end, we study the shape and structure of any lesions (suspicious regions). This includes measuring how sharp the edges are, how big the regions are compared to the whole image, and how regular or irregular their shapes are. This part contributes the last 8 features. Together, these five groups give us a complete 75-feature profile for each image. To better understand how the extracted features contribute to HPV classification, we conducted several analyses on feature behavior, importance, and separability. Using principal component analysis (PCA) and t-SNE, we first began by visualizing

the feature space. These embeddings showed that there was a partial grouping of samples that were positive for HPV and samples that were negative for HPV. This indicates that the recovered features contain information that is discriminatory. An examination of variance using principal component analysis (PCA) found that more than 95% of the variation could be kept with only 15 components, demonstrating that the features were quite compact. We evaluated the mean absolute differences between HPV classes in order to evaluate the diagnostic capability of a number of different feature groups, respectively. Texture was the least discriminative of the three characteristics, whereas morphology and color characteristics were the most distinct.

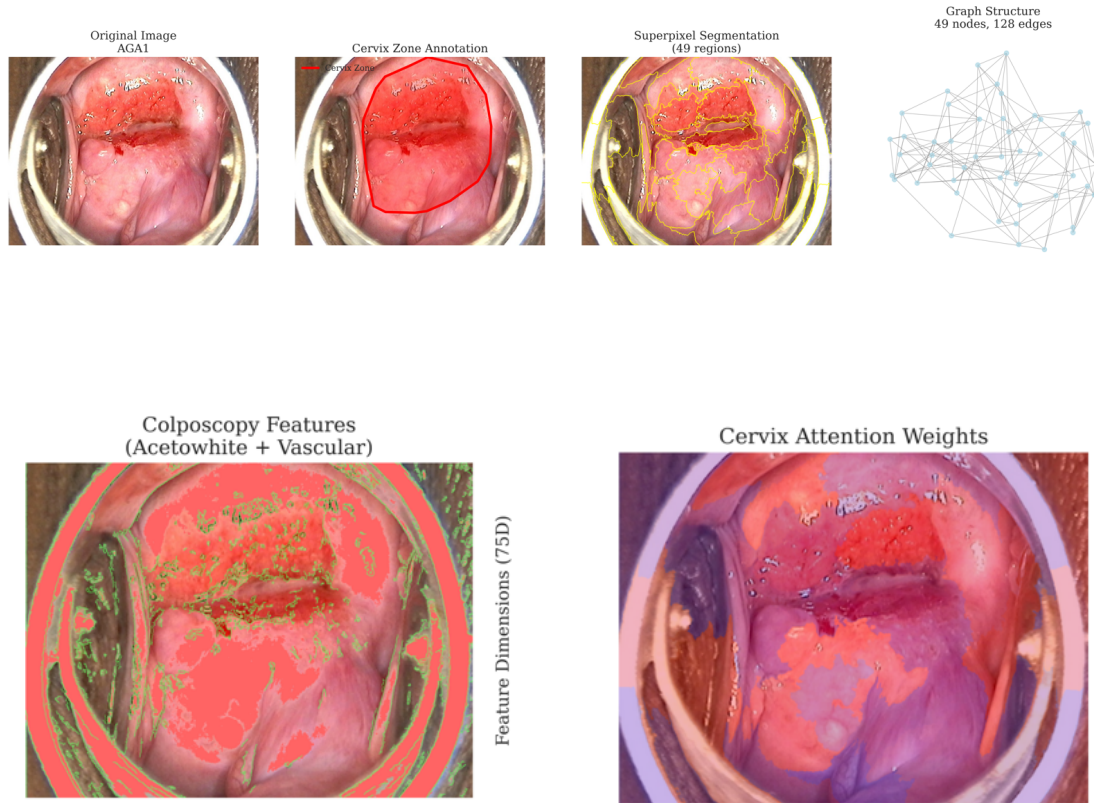


Figure 3.5: Feature Extraction from Colposcopy Images

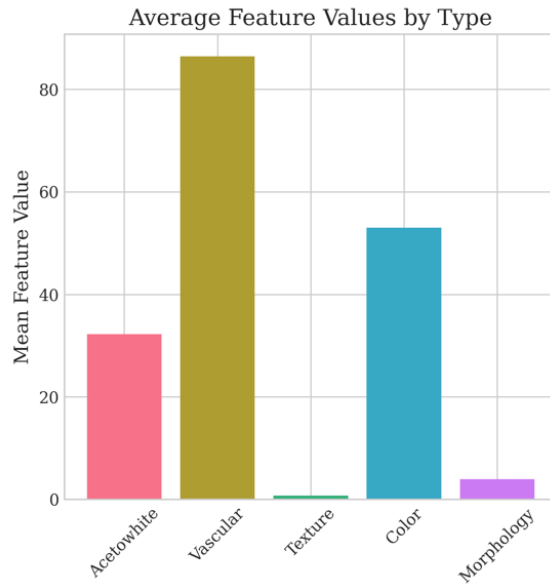


Figure 3.6: Average Feature Values by Type Distribution

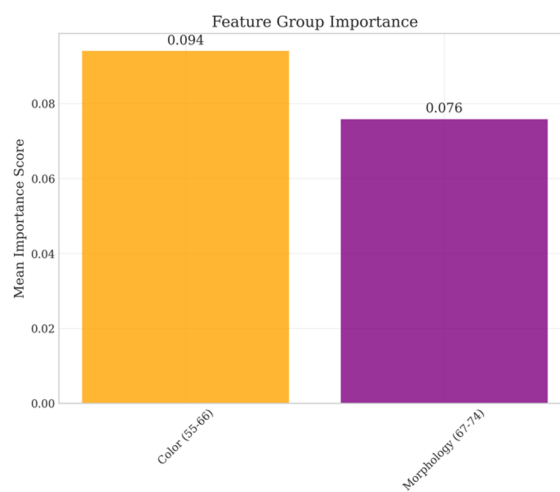
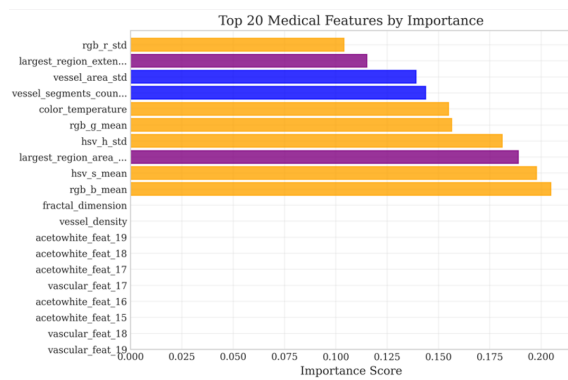


Figure 3.7: Medical Features and Feature Group by Importance

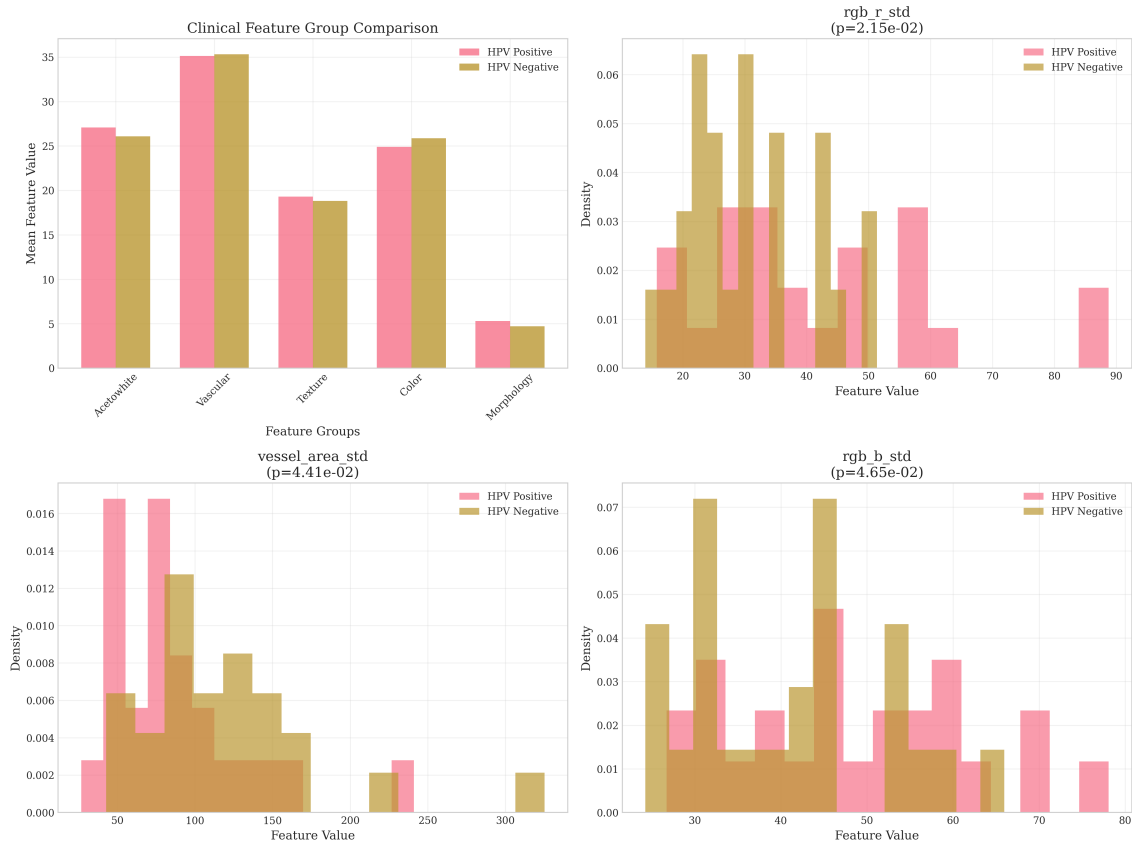


Figure 3.8: Comparative Analysis of Clinical Feature Distributions for HPV Status

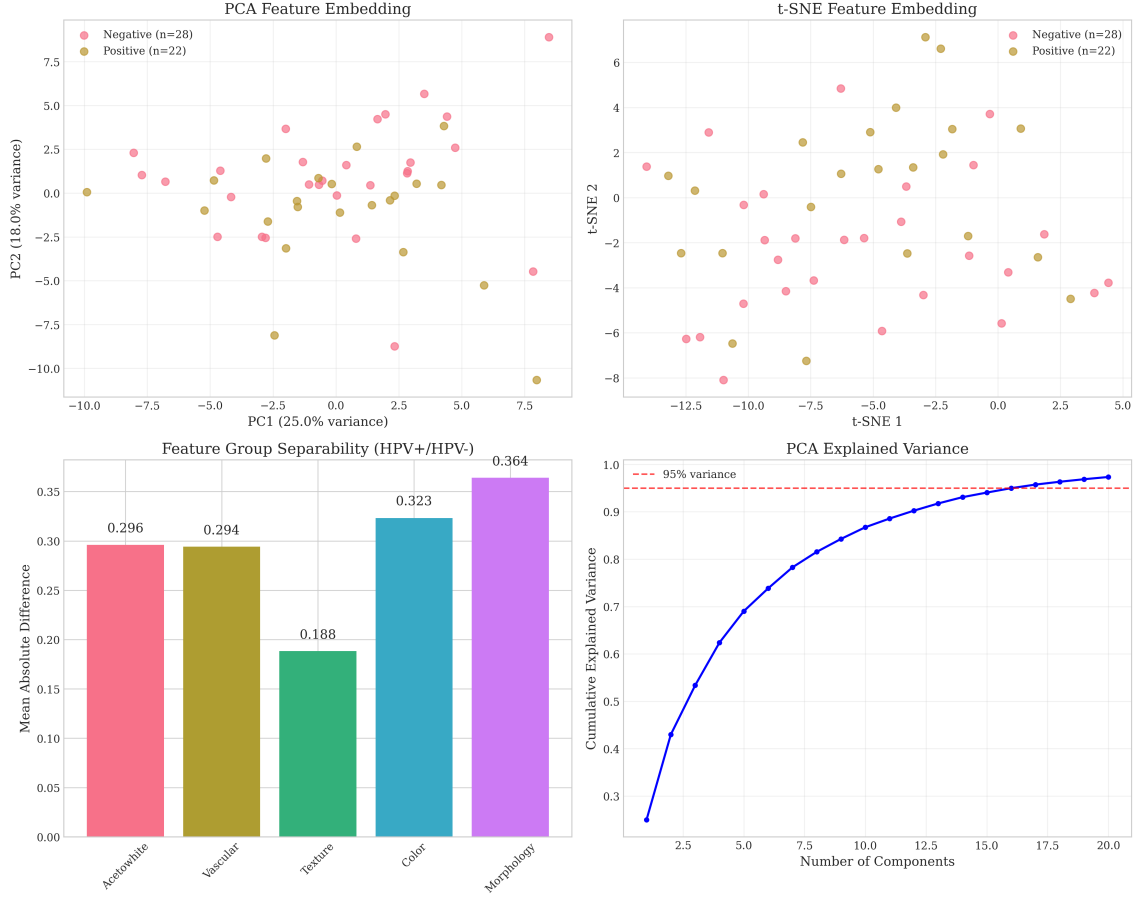


Figure 3.9: PCA and t-SNE Feature Embedding and Feature Group Separability

3.3.6 Graph Construction

To apply these features more effectively in a structured format, we integrated them into a graph-based representation of each image. Each image is divided into small regions called superpixels using the SLIC method, covering the entire image. Using annotation data, we then calculate how much each superpixel overlaps with the cervix zone. These selected superpixels form the Region of Interest (ROI), allowing the model to focus only on the clinically important part of the image. From each of these selected superpixels, a set of 75 features is extracted to represent meaningful medical and visual information. While features are extracted for the entire image, each region’s features are weighted based on how much it overlaps with the cervix zone. We give each region a weight (attention value) based on this overlap. This makes the superpixels inside the cervix stand out more, while the ones outside are down-weighted. This enables the model to learn from all regions, but prioritize the most relevant areas that have the most influence. In the end, every selected region had its own set of features and a score showing how important it is for diagnosing HPV. Then, these superpixels of the annotated part became nodes in a graph. Nodes are connected based on spatial proximity, forming edges in the graph. This graph layout not only shows useful visual information, but it also shows how these areas are connected spatially. This process was done for every image in the dataset. By focusing on the cervix area and organizing the regions into graphs, the final data became more useful for training models that aim to detect HPV and related cervical

conditions accurately.

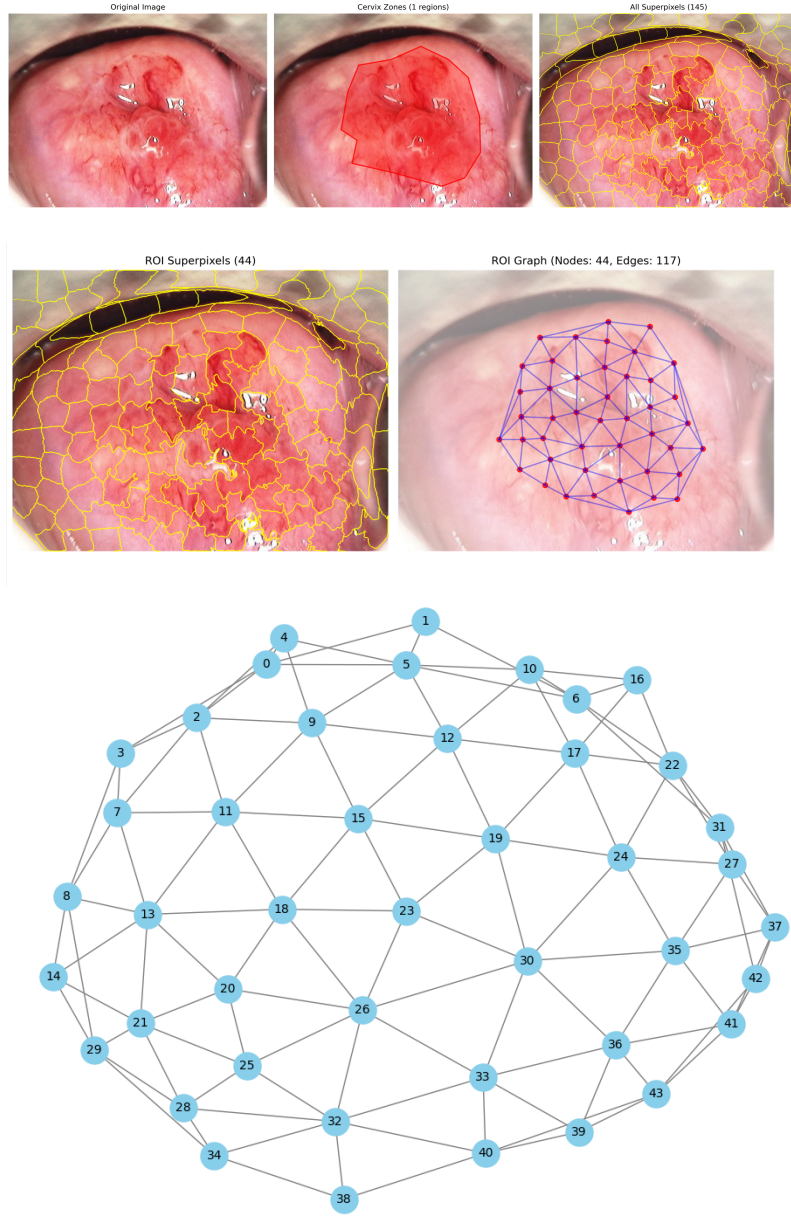


Figure 3.10: Graph construction from original image

3.4 Selected Models Description

In our study on the early detection of cervical cancer via colposcopy images, we specifically chose and implemented certain graph neural network (GNN) models to effectively analyze the complex visual characteristics that are found in cervical tissues. Our selection of models was based on the need to capture different aspects of the data while considering the unique challenges of medical image analysis. We used five different models together to make the most of the capabilities of different GNN architectures.

3.4.1 GraphMAE2

The first primary model we used is GraphMAE2. GraphMAE2 is a self-supervised learning model for graph data that learns meaningful node representations by masking node features and reconstructing them using graph neural networks. Moreover, the encoder in GraphMAE2 is a stack of GAT layers enhanced with residual connections and layer normalization, processing the masked graph to produce rich contextual embeddings, while the decoder is a multi-layer perceptron (MLP) with ELU activations and skip connections that reconstructs the original node features from these embeddings. This self-supervised approach allows the models to acquire robust feature representations in detecting the prominent zone [33].

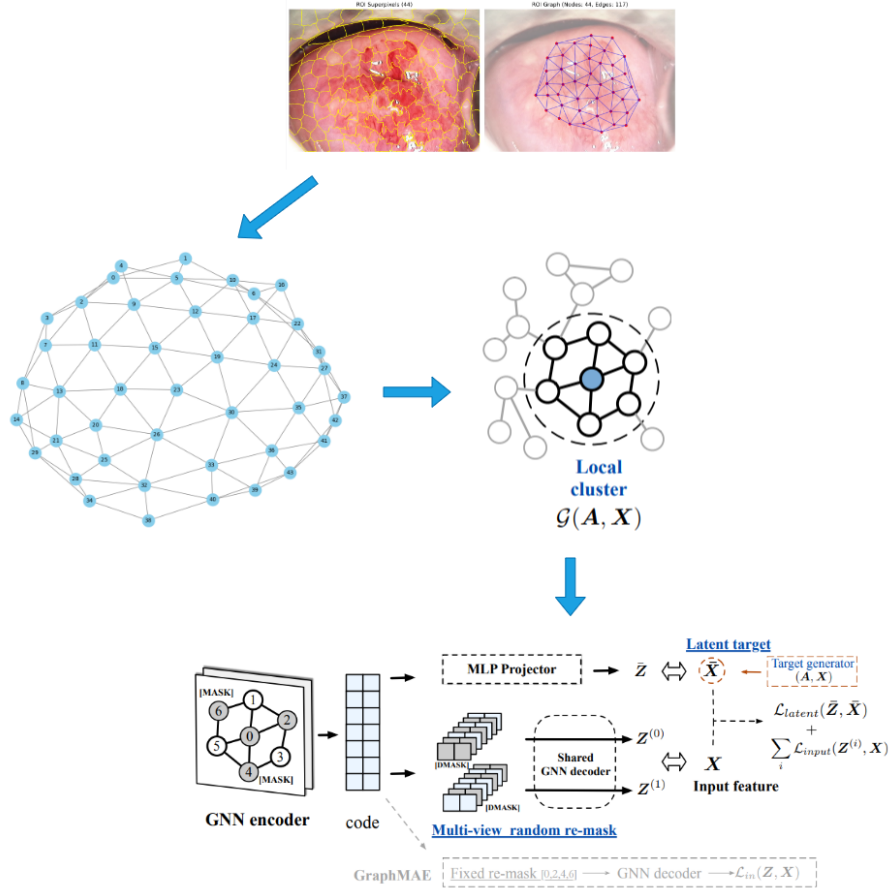


Figure 3.11: GraphMAE2 Model Stucture

3.4.2 Graph Attention Network (GAT)

Graph Attention Network (GAT) is the next model we used. GAT excels at selectively focusing on the most important relationships between nodes through its multi-head attention mechanism, which is particularly valuable for colposcopy analysis, where certain regions may carry more diagnostic significance. Here, the input is all the node features, and those features transform into a high-dimensional feature. Here the self-attention mechanism represents the importance of a node to another node [46].

Graph Attention Networks

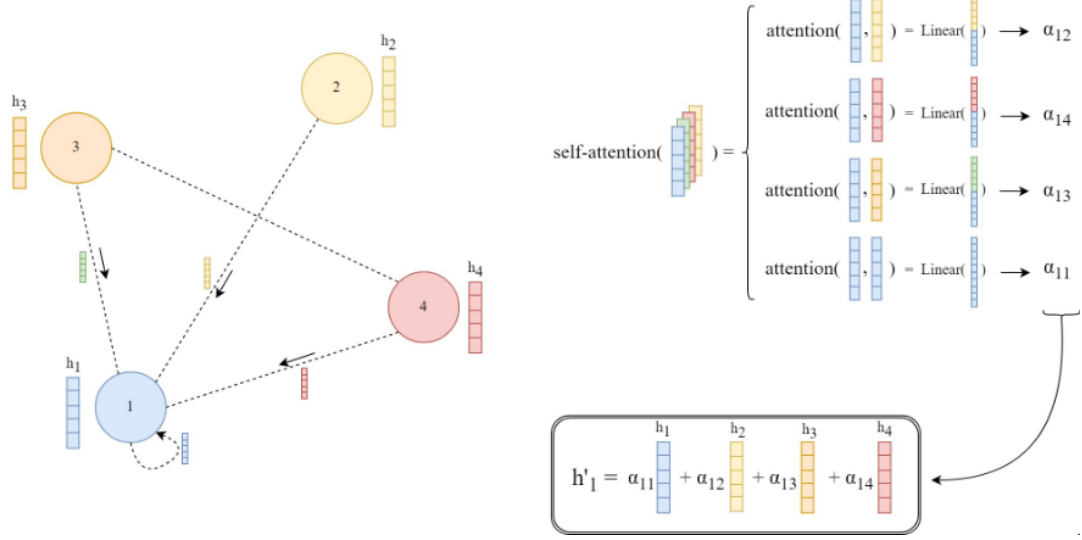


Figure 3.12: Graph Attention Network model Structure [39]

3.4.3 Graph Isomorphism Network (GIN)

We further chose the model Graph Isomorphism Network (GIN), which is known for its strong ability to distinguish between different graph structures. GIN excels at detecting subtle structural differences between normal and abnormal tissue graphs, making it particularly suitable for identifying early-stage abnormalities that might be missed by other models. The theoretical properties of GIN ensure that it can distinguish between different graph structures with high sensitivity, which is crucial for detecting important small changes in tissue architecture [9].

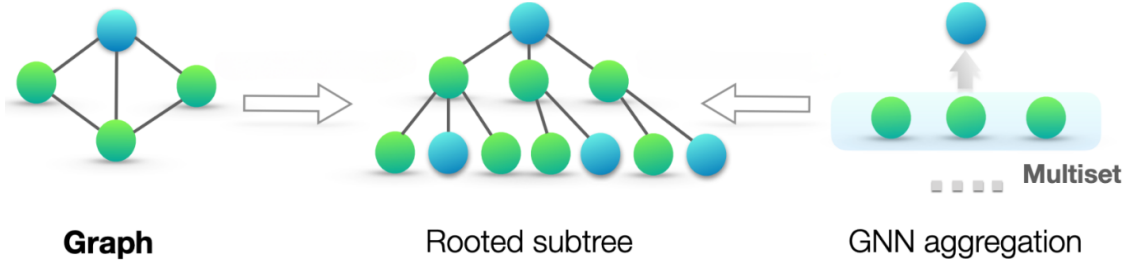


Figure 3.13: Graph Isomorphism Network Model Structure [9]

3.4.4 Graph Convolutional Networks (GCN)

Lastly, we choose Graph Convolutional Networks (GCN). GCN are the type of neural network for working with graph-structured data. GCN operate by allowing every node on the graph to update its representation by collecting the information from the surrounding neighbor nodes. By this process, the network learns to determine the structure of the graph and the features of the individual nodes. GCNs consist of various layers where every layer specializes in the combining and transforming of features of one node and the neighboring nodes. In the GCN architecture, the first layer employs a dimension-matching projection when input and hidden dimen-

sions differ, ensuring consistent tensorial shapes for residual addition operations. Subsequent layers maintain dimensionality consistency, allowing direct residual connections through identity mappings. GCN are typically applied for purposes such as link prediction, node classification, and graph classification. GCN works well and with great efficiency for the majority of applications, especially where relationships between entities which are important [4]

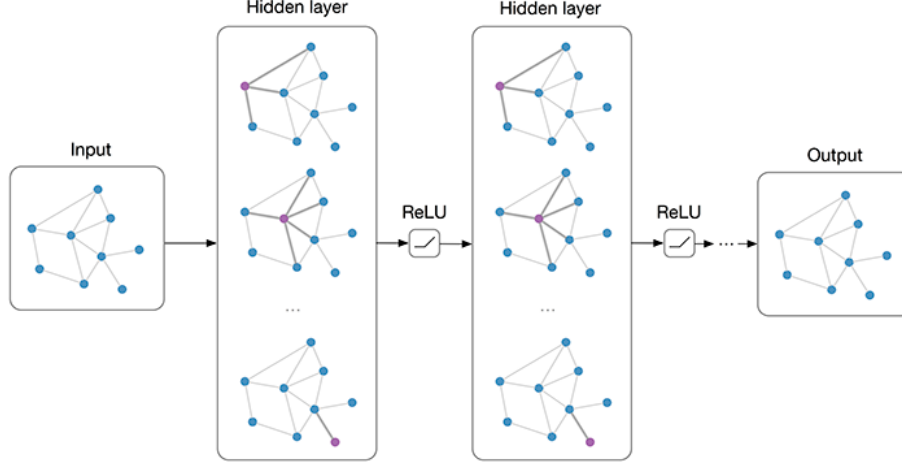


Figure 3.14: Graph Convolutional Network model Structure [3]

We use several graph neural network (GNN) architectures, namely GraphMAE2, GAT, GIN, and GCN, that are capable of dealing with graph-structured data. These differs fundamentally from that of grid-like data typically handled by Convolutional Neural Networks (CNNs). Unlike images, whose pixels are mapped directly to a fixed grid, our data is graph-structured by its very nature: every sample consists of a set of nodes and edges, accompanied by complex features of these nodes and edges representing relationships and interactions between them. Traditional CNNs are optimized for spatially localized, grid-structured data, and thus these are ill-suited for capturing the arbitrary and potentially nonlocal relationships present in graph data. Applying CNNs on such data would require graph structure to be flattened or converted, which would necessarily result in the loss of important relational information.

Graph Convolutional Network (GCN) aggregates features from neighboring nodes, effectively capturing the local structures existing in the graph. Graph Attention Network (GAT) introduces attention mechanisms to weight the importance of neighboring nodes, which allows the model to focus on the most important sections of the graph. The Graph Isomorphism Network (GIN) has theoretical capabilities that match those of the Weisfeiler-Lehman graph isomorphism test, allowing the model to distinguish complex graph topologies. Additionally, GraphMAE2 makes use of masked autoencoding methods to obtain robust representations of the nodes by reconstructing latent node features, thus allowing the model to capture subtle complexities within the graph. In summary, our carefully selected methods of graph neural networks, together with feature engineering that is specialized to the field and sophisticated training methods, make a strong and therapeutically useful system for finding cervical cancer early in colposcopy images. This approach connects the gap between clinical expertise and artificial intelligence, providing a reliable tool for automated screening and decision support in healthcare settings.

Chapter 4

Model Implementation

We designed a comprehensive training methodology to develop a strong and reliable model for classifying cervical images. This methodology emphasizes performance, stability, and interpretability. This process includes a series of decisions on architecture, tactics for pretraining, integration of ensembles, and methods for assessment. These are made to find useful patterns in medical data that is in the form of graphs.

4.1 Proposed GNN Architecture Components

4.1.1 Customized Graph Masked Autoencoder (GraphMAE2)

We have created two different variants of the GraphMAE2 model, which is based on the GraphMAE2 architecture but customized with a cervix-aware attention mechanism. We have used one large variant and one small variant. The large variant uses 512 hidden channels, while the small variant uses 256, allowing us to balance between model complexity and computational efficiency. Both models, configured with 8 attention heads across 3 layers, benefit from a masked autoencoding pre-training strategy. These models use a self-supervised learning strategy where portions of node features are masked and then reconstructed using the surrounding graph context. This approach is particularly helpful in the extraction of structural representations that are meaningful. Additionally, this method does not rely on labeled data, which makes it especially useful in domains with limited annotation. This selective reconstruction target focuses the learning process on developing predictive capabilities from contextual information rather than merely memorizing the input. The GraphMAE2 model employs focal loss with the help of alpha balancing to handle class imbalance. This loss function particularly focuses on hard-to-classify cases and is necessary for managing the under-representation of positive cases in cervical cancer detection.

$$\mathcal{L}_{\text{focal}} = \alpha \cdot (1 - p_t)^\gamma \cdot \mathcal{L}_{\text{CE}} \quad (4.1)$$

Here, $p_t = e^{-\mathcal{L}_{\text{CE}}}$ represents the model's confidence level, $\alpha = 0.25$ is used as a balancing parameter, $\gamma = 2.0$ is used as a focusing parameter.

Moreover, within our framework, GraphMAE2 contributes by normalizing its weights adaptively within the ensemble so it can dynamically control its own influence. This approach secures that the strength of GraphMAE2 is utilized optimally within the

ensemble.

$$w_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (4.2)$$

Here, z_i represents the learnable parameters, and $K = 5$ represents the number of models.

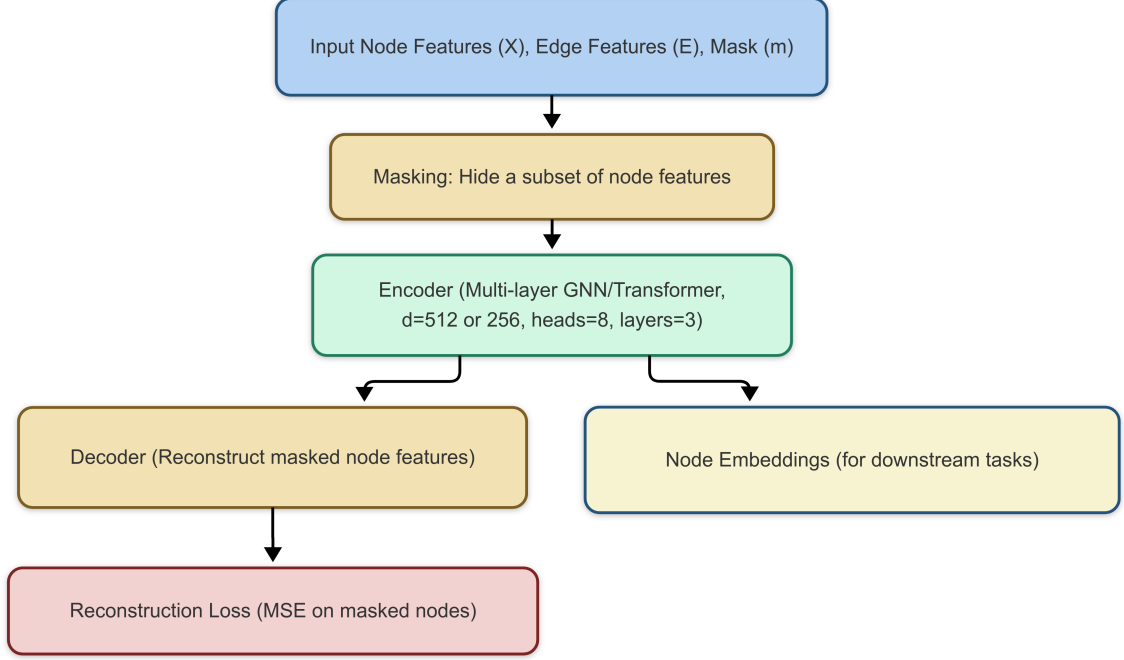


Figure 4.1: Architecture of Customized GraphMAE2 Model

4.1.2 Customized Graph Isomorphism Network (GINModel)

In order to differentiate between various graph architectures, the GINModel was created to be very expressive. Our GIN implementation uses 512 hidden channels and incorporates multi-layer perceptrons (MLPs) with batch normalization for stable training and GELU(Gaussian Error Linear Unit) activations for smooth learning. We used the architecture that incorporates dropout regularization with a rate of 0.5, which randomly turns off neurons in training to reduce dependence between neurons and improve generalization skill. This combination of architectural choices makes the GIN component to effectively distinguish between normal and abnormal cellular graph patterns that might be visually subtle but topologically significant. Moreover, GIN model has a compounded loss function integrating focal loss and label smoothing with an aim of enhancing robustness of the introduced model. Its hybrid mechanism prevents overconfidence and places specific stress on exactly the hard classification instances.

$$\mathcal{L}_{\text{combined}} = \alpha \cdot \mathcal{L}_{\text{focal}} + (1 - \alpha) \cdot \mathcal{L}_{\text{CE-LS}} \quad (4.3)$$

$\alpha = 0.5$ represents the weighting parameter.

Again, in an attempt to make training balanced, GIN model suggests class-balanced sample weights. This corrects class distribution imbalance by offering equal class representation of images.

$$w_c = \frac{1}{N_c} \quad (4.4)$$

Here, N_c represents the number of class inc.

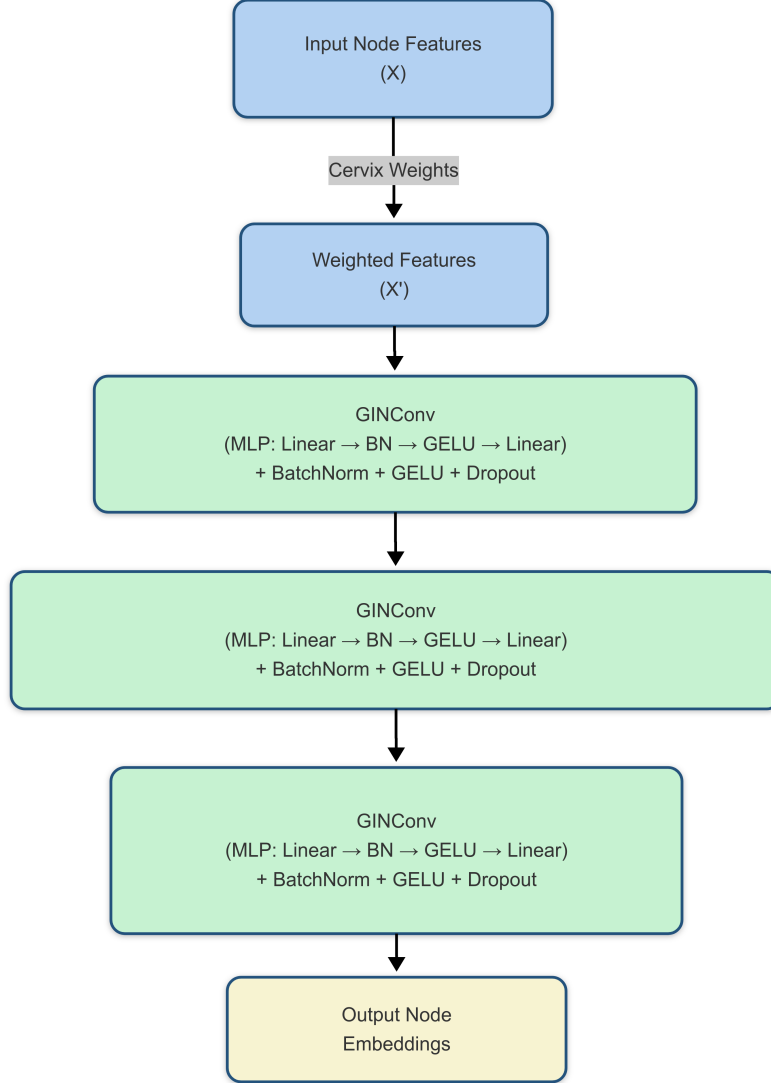


Figure 4.2: Architecture of Customized GIN Model

4.1.3 Customized Graph Convolutional Network (GCNModel)

Our customized GCN provides a strong baseline for graph data and capitalizes on effective message-passing mechanisms. We customized it with 512 hidden channels and residual connections. This model is based on the normal GCN architecture, but it adds residual connections. Our customized model improves gradient flow during backpropagation and allows for deeper architectural configurations without performance degradation. An additional enhancement of our customized GCN implementation is cervix-aware node weighting, which we can call weighted multia-modal

fusion, which allows prioritization of clinically relevant areas of high importance. The fusion is defined as:

$$h = \sum_{i=1}^K w_i \cdot h_i \quad (4.5)$$

Here, w_i represents the normalized weight for model i , and $K = 5$ represents the number of models.

This weighting mechanism combines representations from multiple architectures and applied to node features. The weights are derived from knowledge of cervical anatomy and pathology that is particular to the field. Moreover, GCN applies stochastic depth regularization where $r \sim \text{Uniform}(0, 1)$, by which this model can prevent co-adaption, which can be happened by randomly dropping model contributions during training. Regularization is defined as :

$$h_i = \begin{cases} F_{\text{dropout}}(h_i, p = 0.2), & \text{if } r > 0.1 \\ 0, & \text{otherwise} \end{cases} \quad (4.6)$$

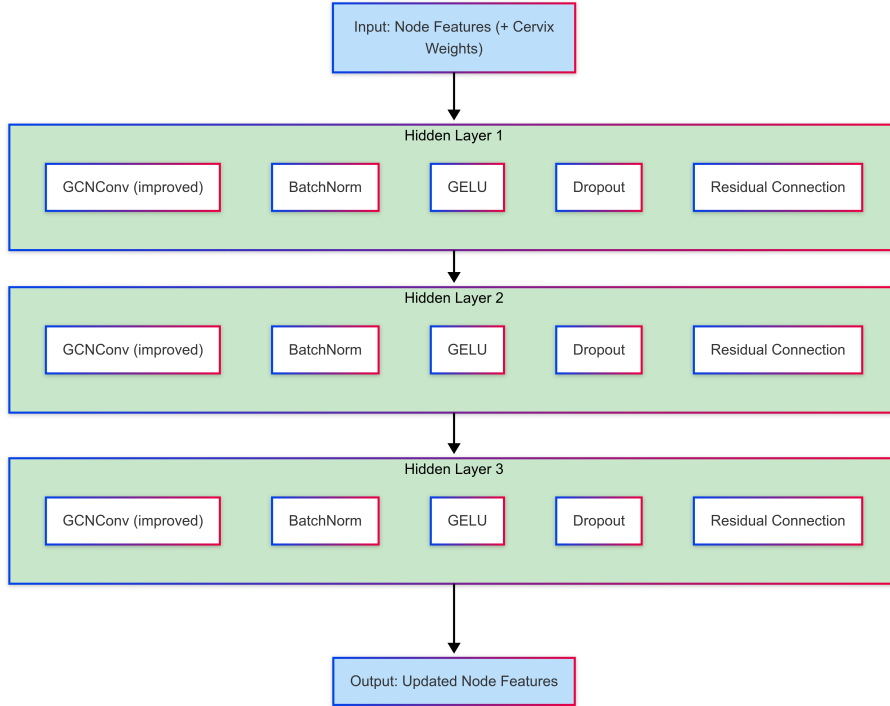


Figure 4.3: Architecture of Customized GCN Model

4.1.4 Customized Graph Attention Network (GATModel)

Our Enhanced Graph Attention Network utilizes multi-head self-attention to selectively weight node relationships according to learned feature interactions. The model is able to choose to emphasize clinically informative graph patterns while reducing the influence of less informative areas in this manner. Our implementation uses 512 hidden channels and 8 attention heads and incorporates improvements such as edge features and batch normalization for improved stability during training.

This technique enhances model performance as diverse approaches to the node relationships are captured in parallel. Each attention head computes coefficients that specify the degree to which adjacent nodes affect the update of the target node’s features. The network architecture transforms the features of interest over several layers and retains the multi-head attention mechanism until the last layer, which utilizes one head to generate a consolidated output representation. In addition, our implementation incorporates cervix-aware attention scaling at the first layer, which allows attention distribution to be influenced by relevant domain knowledge at the first step. Additionally, masked attention is also injected into that model, which works by dropping structural features. The GAT model further enhances calibration through temperature scaling, where $T = 1.5$, to make the probability distribution and confidence scores more reliable.

Calibrated probabilities formula:

$$p_i = \frac{e^{z_i/T}}{\sum_{j=1}^C e^{z_j/T}} \quad (4.7)$$

Here, z_i is the logit for class i ,
 $T = 1.5$ is the temperature parameter, and
 C represents the number of classes

Again, multiple-scale graph pooling is used by the GAT model to represent distinct aspects of graph structure. This approach offers a thorough embedding for tasks of classification across the graph level.

The pooling strategy is defined as:

$$h_{\text{pooled}} = [h_{\text{mean}} || h_{\text{max}} || h_{\text{add}}] \quad (4.8)$$

Here, h_{mean} , h_{max} , and h_{add} represents the pooling operations.

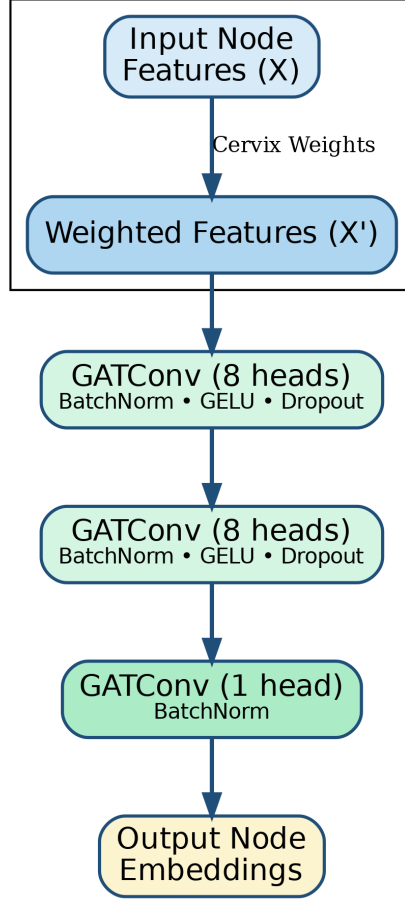


Figure 4.4: Architecture of Customized GAT Model

4.2 Loss Functions and Classification Components

4.2.1 Focal Loss Implementation

We implemented Focal Loss to manage class imbalance inherent in medical datasets. Focal Loss emphasizes the samples, which are hard to classify, by adjusting the loss contribution based on prediction confidence. The Focal Loss modifies the standard cross-entropy loss by introducing a modulating factor

$$(1 - p_t)^\gamma \quad (4.9)$$

Here, p_t is the model's estimated probability for the right class, and the γ is a sharpening parameter that determines the rate at which the simple examples are down-weighted. In our implementation, we determined $\gamma=2.0$ to provide optimal focus on challenging samples. Additionally, an alpha-balancing factor of 0.25 was incorporated to address the class imbalance directly by assigning different weights to positive and negative examples. This dual approach, which focuses on difficult samples while balancing class contributions, proved effective in navigating the complex loss landscape which is presented by imbalanced medical data. In our final implementation, the Focal Loss is combined with label smoothing (0.2) to further enhance model calibration, and this prevents overconfidence. This combined approach reduces the risk of overconfident misclassification, which is a critical consideration

in medical diagnostic applications where both accuracy and appropriate uncertainty estimation are essential.

4.2.2 Ensemble Classifier

The feature extraction process concludes with the use of a powerful ensemble classifier. This classifier enables the mapping of node-level embeddings to graph-level predictions. The classifier uses a complex multi-scale pooling approach that considers different views of the structure of the graph. Our suggested multi-scale pooling approach entails the use of three different aggregation functions. These include mean pooling, maximum pooling, and additive pooling. Mean pooling captures the average features dispersed among all nodes. Maximum pooling emphasizes the most prominent features in the data. Additive pooling preserves the cumulative information found across the graph. The combination of these three pooling methods produces a complete graph-level representation that captures both typical and exceptional features of the input graph.

The classifier’s architecture, following the pooling process, utilizes a hierarchical method of dimension reduction, progressively reducing the representation from the concatenation pooling dimension to an intermediate form in the hidden dimension, then to half of that dimension, and finally to the classification dimension of the output. In order to facilitate improved generalizability, we utilized the full regularization strategy, incorporating complementary methods: batch normalization stabilizes activations within mini-batches, layer normalization normalizes activations between features, standard dropout (0.5) randomly deactivates neurons and spatial dropout (0.4) deactivates full channels. This multi-faceted approach to regularization effectively prevents overfitting despite the limited size of medical datasets.

A distinctive feature of our classifier is its dedicated uncertainty estimation branch, which produces a scalar uncertainty score for each prediction. This branch consists of a separate neural network that processes the pooled graph representation to estimate the model’s confidence in its prediction. The uncertainty assessments are calibrated with a sigmoid activation function to output values between 0 and 1, where high values indicate high uncertainty. This explicit representation of uncertainty provides valuable clinical decision support by identifying cases that may require additional expert review.

4.3 Core Ensemble Architecture

4.3.1 Model Integration

Our ensemble architecture is based on five complementary graph neural networks, which apply a different inductive bias to each model to improve the performance of the system as a whole. The five combined models include two GraphMAE2 variants with different representation capacities (512 and 256 hidden dimensions), a customized GAT (512 hidden dimensions), a customized GIN (512 hidden dimensions), and a customized GCN (512 hidden dimensions). The models process the same input graph but use different operations and structural inductive biases and (512 hidden dimensions).

In order to adaptively combine these different models, we utilized a learnable weighting mechanism. The model is initialized with some bias in favor of the GCN component (weight 1.3 versus 1.0 for all of the others) based on preliminary experiments, but the weights are learned in the course of training through backpropagation. The weights are softmax-normalized to make them add to one in order to compute an appropriately defined weighted average of the component representations.

This adaptive weighting mechanism allows the ensemble to be directed to the effective architectural components for each specific dataset or task, providing an architecture search in the process of training. It was particularly effective in medical image analysis, wherein multiple architectural inductive biases can encapsulate multiple aspects of the pathology.

4.3.2 Pre-Training Methodology

Prior to supervised learning, the GraphMAE2 models undergo pretraining using masked autoencoding. Approximately 15% of node features are randomly masked, and the models are trained to reconstruct these features. This enhances their ability to generalize from structure alone. Gradient clipping is applied to stabilize the training process, and reconstruction loss is monitored across epochs. A self-supervised pre-training technique was employed for the GraphMAE2 architecture to enhance the learning capacity and generalization of our graph-based models. The pretraining phase enables the model to acquire strong and significant feature representations from graph-structured colposcopy data. In the pre training phase, a masked autoencoding technique is employed for each epoch, where 15% of the node features are randomly chosen and masked, effectively setting their values to zero or concealing them from the model. The encoder of Graphmae2 effectively processes the partially masked graph, acquiring the ability to gather and spread information from the visible nodes. On the other hand, the decoder endeavors to reconstruct the original features of the masked nodes, presenting a challenge for the model to extract missing information from the context of the remaining graph. This includes every graph present in the dataset. The targeted reconstruction approach compels the model to cultivate a strong internal representation of the graph structure and the distributions of node features. The quantification of this learning is achieved through pretraining loss, characterized as the mean squared error between the predicted and actual features of the masked nodes, with an emphasis on clinically significant regions such as the cervix zone. Both variants of GraphMAE2, large and small, undergo simultaneous training utilizing a combined loss function that aggregates their respective reconstruction losses. This collaborative training enables the models to cultivate complementary representations that can subsequently be utilized in the ensemble.

In order to enhance the stability of the pretraining process, we applied gradient clipping, setting a maximum norm of 1.0. Furthermore, the system identifies and addresses numerical instabilities, including NaN or infinite loss values, and it is done by omitting the impacted batches. The implementation of these reliability measures guarantees consistent convergence, regardless of the intricate nature of the pretraining. A slightly faster learning rate (1.5 times the guided learning rate) was used in the pretraining phase to make studying the representation space easier. This method lets the model get better at extracting general features before it is fine-tuned for the classification. The pretraining phase spans over 20 epochs, during which the

model’s performance is assessed and preserved at its optimal state for subsequent fine-tuning. To handle multiple graphs at once, we have implemented optimized data loading and batching techniques to maximize the use of computational resources. By focusing on reconstructing masked features, the model builds a strong understanding of the dataset. In the end, this helps to guide the training in the future. By doing things this way, the model starts off strong, which lets it learn from small amounts of labeled data. This helps in adjusting to new situations, and finally makes accurate predictions.

4.3.3 Forward Propagation Logic

The forward propagation process integrates outputs from all the components of the model in an advanced sequence of transformation. This process is initiated by parallel feature extraction of all of the model components, followed by dimension harmonization, normalization, regularization, and adaptive fusion. First of all, each of the component models processes the input graph independently to generate its representation of features. Because of architectural differences that could potentially use different dimension sizes for these representations, dynamic dimension detection and initialization of the projection layers were utilized. The projection layers project each model’s representation to a common hidden dimension (which is fixed during initialization) such that the diverse representations could be reasonably combined. Following the process of the projection, batch normalization is applied to each individual model’s output. This normalization stabilizes the feature distribution without allowing any individual model to overpower the entire ensemble due to scale variations. The normalized features are regularized model-specifically in order to improve generalization capabilities. During training, we employ a stochastic depth mechanism with a drop probability of 0.1. This technique randomly deactivates entire model contributions with a 10% probability, replacing them with zero tensors. When a model is not completely deactivated, standard dropout (0.2) is applied to its output. This dual-level dropout approach, at both the model and feature levels, creates a robust ensemble that does not overly depend on any single architectural component.

The regularized and normalized representations are then combined utilizing learned ensemble weights. The weights undergo softmax normalization to ensure they sum to one to create an actual weighted average. The combined representation is then processed through the classifier in order to produce logits and uncertainty estimates. During inference, the logits are temperature-scaled with $T=1.5$ in order to improve probability calibration. The scaling reduces the confidence of the model in its prediction to avoid overconfidence often observed in neural networks.

4.4 Training Framework

4.4.1 Cross-Validation Setup

To obtain stable estimates of model performance as well as to guide model selection, we used a stratified k-fold cross-validation approach with $k=5$. This method is especially valuable in medical research with limited sample sizes in that it allows us to use all of the available data to the best possible advantage while it is

still able to provide good performance estimation. Stratification ensures that every cross-validation fold has the same class distribution as the entire data set. It is most important in imbalanced data that the minority class does not necessarily end up becoming underrepresented in any fold. By maintaining a consistent class distribution for all folds, it is far easier to make comparisons of model performance on different splits that are fair. During each fold, we further addressed class imbalance by using class-balanced sampling of training data. In this technique, each sample’s probability of selection is assigned in an inverse proportion to its class frequency. That is, minority class samples are assigned a higher chance of selection while training. This process avoids the model becoming skewed in favor of the most common class and helps it learn appropriately from the less common but clinically relevant examples. Performance metrics scores such as accuracy, precision, recall, F1 score, and estimates of uncertainty were monitored continuously in each fold to facilitate detailed tracking of performance as well as between-split comparisons.

4.4.2 Model Training

Our three-step training process involves pretraining, main training, and fine-tuning. This progressive approach allows the model to develop from general representational capabilities to specialization in performing specific tasks in an organized manner.

The pretraining phase, which is conducted over 20 epochs, is specific to the Graph-MAE2 components. It utilizes the AdamW optimizer with an increased learning rate 1.5 times higher compared to that in the main train phase ($3e-4$ versus $2e-4$) with the same weight decay value of $2e-4$ used for regularization. The higher learning rate is used to enable further exploration of the representation space in this self-supervised learning stage.

Following pretraining, the full ensemble model with all five architectural components is added in the full training phase. It uses the AdamW optimizer with a learning rate of $2e-4$ and weight decay of $2e-4$. For enhancing training stability and convergence, we implemented a cosine annealing learning rate scheduling with a minimum learning rate of $1e-5$. This learning rate scheduling gradually reduces the learning rate linearly along a cosine curve progressively during the training process so that the model learns to converge to sharper minima during training.

The initial training process is based on using early stopping with a validation F1 score of patience 20 epochs. This mechanism of preventing overfitting stops the train when the validation set performance does not improve any further. The best-performing model state is saved for subsequent fine-tuning as well as for testing purposes.

After the main training period is over, the best-performing model then goes through the process of fine-tuning with an extremely low learning rate of $2e-5$ so that there can be more precise parameter adjustments. This phase employs a cosine annealing schedule with $1e-6$ as its lowest learning rate, but with an increased patience of 30 epochs in case of early stopping. This conservative approach will enable the model to make minor adjustments in its parameters without disturbing the representation

achieved in the initial phase of training.

4.4.3 Training Loop Implementation

Our training loop design unifies a number of advanced methods in an attempt to amplify efficiency and reliability. Some of them include gradient accumulation, loss functions that are adapted, and wide-ranging metric monitoring as well. In order to scale up the batch size appropriately without scaling up memory requirements proportionally, we employed gradient accumulation with an accumulation factor of 2. Consequently, this technique divides the loss by the accumulation factor before backpropagation and accumulates gradients across multiple forward-backward passes before applying an optimizer update. Furthermore, this approach doubles the batch size from 16 to 32 samples. This helps in improving gradient estimate quality while maintaining memory efficiency also.

Again, for graphs with available cervix weighting information, weights are applied during forward propagation to emphasize clinically relevant regions. This domain-specific attention mechanism directs the model’s focus toward anatomically significant structures, which helps to enhance both learning efficiency and interoperability. Our loss computation combines two complementary objectives mainly Focal Loss for addressing class imbalance and cross-entropy with label smoothing (0.2) for improved calibration. So, these components are weighted with a balancing factor =0.5, providing equal importance to both loss terms. This combined objective function simultaneously addresses class imbalance, along with hard example emphasis, and confidence calibration.

Gradient clipping is also applied with a maximum norm of 3.0 during the main training phase to maintain numerical stability during backpropagation. This helps in preventing gradient explosions that can destabilize the training process, particularly seen in deeply structured networks with diverse components. During training, we keep track of many performance measure such as accuracy, F1 score, precision, recall, loss, and uncertainty for both the training and validation sets.

4.4.4 Fine-Tuning Phase

Our training method uses a two-step early stopping approach along with model preservation and fine-tuning to get the best results in the end stage. This method effectively harmonizes the exploration phase in the early training stages with the exploitation phase during fine-tuning. The performance of the model on the validation set is consistently monitored during the primary training phase. The F1 score is examined in greater detail. This metric effectively considers both accuracy and recall. In medical diagnostic tasks, metrics such as recall, F1 score, precision, and accuracy are essential due to the substantial consequences that false positives and false negatives can impose. To prevent overfitting to the training dataset, the primary training phase ends when it is observed that the validation F1 score does not improve over 20 consecutive epochs. The model state that got the best F1 score for validation is kept and used as a starting point for the fine-tuning process. The fine-tuning phase utilizes a notably lower learning rate ($2e-5$) to facilitate more cautious parameter adjustments. This cautious method keeps the learned representations from changing in big ways while still letting subtle changes happen that improves performance.

Gradient clipping uses a lower maximum norm (1.0 vs. 3.0 in main training) during fine-tuning to further limit the size of parameter changes. The fine-tuning process will endure until the validation F1 score is stalled for 30 epochs consecutively. This phase has greater patience than the initial training phase. Furthermore, the reduced learning rate allows the fine-tuning process to implement slight modifications that are not promptly obvious. The finetuning phase has a greater opportunity to do this due to the higher patience. When fine-tuning is performed, the model states from the two steps (main training and fine-tuning) are compared based on the validation F1 scores. The top model is selected for final testing, which ensures that the fine-tuning process is only accepted if it actually improves things.

Chapter 5

Performance Evaluation

5.1 Experimental Setup

Our experiments were performed at the VSCODE code editor. For processing, an Intel Core i7 14700K CPU, with its huge processing capabilities. For heavy processing and computation, an NVIDIA RTX4080 Super GPU accompanies the system with 16GB of VRAM, ideal for computation and graphical processing requirements. For storing, 1TB SSD takes care, with quick access to files and information. For efficient multitask performance, 64GB RAM comes with the system, with a smooth performance and multi task ability. In general, such a configuration appears to represent a well-maintained and flexible configuration for our experimental environment.

Component	Specification
CPU	Intel Core i7-14700K
Graphics	NVIDIA RTX 4080 Super
Storage	1TB NVMe SSD
RAM	64 GB

Table 5.1: Hardware Specifications

5.2 Evaluation Metrics

To evaluate the performance of our model post-training, it is crucial to use suitable measures. In a dataset with a balanced class distribution, measurements such as total classification accuracy may be enough. Nonetheless, for imbalanced datasets, such as ours, a confusion matrix is very beneficial. A confusion matrix is a simple report of correct and incorrect predictions, in terms of actual and predicted classes. Most important parts in a confusion matrix include:

- **True Positives (TP):** Model predicted a positive class correctly, in agreement with actual positive label.
- **True Negatives (TN):** Model predicted a correct negative class, in agreement with actual label for a negative case.

- **False Positives (FP):** Model wrongly predicted a positive class, when actually its label is not positive. They can alternatively be considered Type I errors.
- **False Negatives (FN):** Model wrongly predicted a negative class, when actually its label is positive. They can alternatively be considered Type II errors.

5.2.1 Accuracy

An estimation of a proportion of correct predictions (negative and positive both) out of a model forecast count. Accuracy reflects the percentage of cases that the model predicted accurately. It is a significant indicator for assessing the effectiveness of classification.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i) \quad (5.1)$$

5.2.2 Precision

Precision measures a model's accuracy in terms of predicting positive ones out of all predicted positive ones. Precision mirrors dependability in positive model prediction and, therefore, in scenarios in which consequences of a wrong positive have high repercussions.

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \quad (5.2)$$

$$\text{Macro-Precision} = \frac{1}{C} \sum_{k=1}^C \text{Precision}_k \quad (5.3)$$

5.2.3 Recall

Recall, alternatively termed as sensitivity, or positive predictive value, estimates a model's ability for predicting all actual positive occurrences. Recall is most important when failure in positive examples (the cases of wrong negatives) entails considerable repercussions.

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \quad (5.4)$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{k=1}^C \text{Recall}_k \quad (5.5)$$

5.2.4 F1 Score

The F1 score gives a harmonic mean of precision and recall, a balanced measure, both wrong positive and wrong negative cases in consideration. When working with an imbalanced dataset, it helps with both precision and recall.

$$F1_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \quad (5.6)$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{k=1}^C F1_k \quad (5.7)$$

5.2.5 Uncertainty

One important feature of the ensemble model that we propose here is that the model can quantify predictive uncertainty for every sample. We estimated uncertainty in our work as average predictive uncertainty across the ensemble models for every input, as implemented in the evaluation pipeline. This approach provides a direct measure of the model’s confidence in its predictions, and the larger the uncertainty, the less reliable outputs. The uncertainty is estimated by a dedicated head in the classifier. After pooling and concatenating node embeddings, the vector is sent through a small MLP ending with a sigmoid activation, producing a value in $[0,1]$ for every sample. This value shows the model’s estimated uncertainty (greater value indicates lower confidence).

$$u = \sigma \left(\mathbf{w}_2^\top \cdot \text{Dropout} \left(\phi \left(\mathbf{W}1\mathbf{h}_{\text{pooled}} + \mathbf{b}_1 \right) \right) + b_2 \right) \quad (5.8)$$

Providing not only predictions but also the levels of uncertainty attached to them, it enables safer and better informed decision-making, as needed in high-stakes applications such as the analysis of colposcopy images.

5.3 Comparative Analysis

5.3.1 Performance Analysis of GraphMAE2 Model

GraphMAE2 was the first model we incorporated into the construction of the ensemble graph neural network. Additionally, we assessed the GraphMAE2 model’s performance using our dataset. We experimented with our dataset of 1279 samples, each represented as a graph with 75-dimensional node characteristics, with the goal to assess the effectiveness of the GraphMAE2 model. GraphMAE2 employs a masked autoencoding pretraining technique that is tailored to graph-structured data utilizing Graph Attention Networks (GATs). In the pretraining stage, we mask 15% of the node features. The GraphMAE2 encoder is composed of stacked multi-head GAT layers with layer normalization, ELU activation, and dropout, facilitating stable and expressive node-level representation learning. On the other hand, The GraphMAE2 model’s decoder is a lightweight MLP that allows self-supervised pre-training by reconstructing the masked node characteristics. The average squared loss calculated solely over the masked nodes is used later to train the model. Moreover, Substantial converging and feature learning were demonstrated by the reconstruction loss, which steadily dropped from 1.01 to roughly 0.74 throughout pretraining before coming to a early stopping. Additionally, the pre-trained encoder is fine-tuned with a graph-level classifier that pools global mean and max pooling of the encoded node representations. In order to counter the extreme class imbalance of the dataset, we trained the classifier using Focal Loss and empirically determined class

weights. As a result, The GraphMAE2 model achieved a validation F1-score of 69%, demonstrating itself to be an effective base model. Though an ensemble model was subsequently used to further maximize performance, GraphMAE2 provides a strong and interpretable baseline, and it establishes the benefits of masked autoencoding and GAT-based encoding for biomedical graph representation learning.

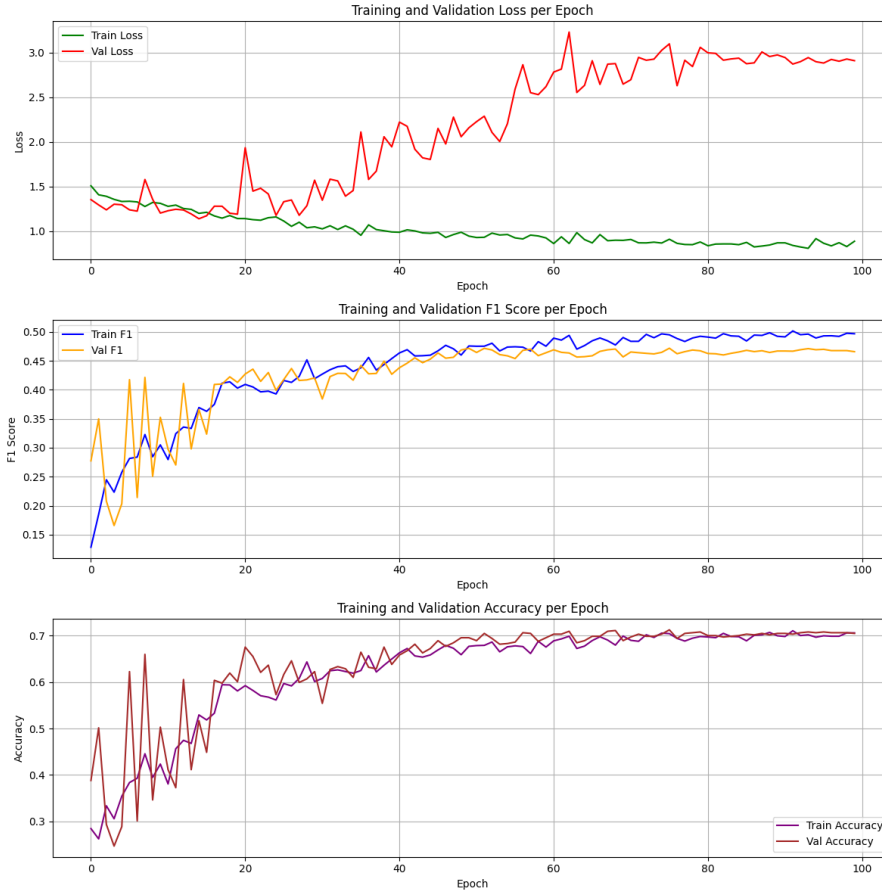


Figure 5.1: Individual Customized GraphMAE2

5.3.2 Performance Analysis of the Ensemble Graph Neural Network

The predictive accuracy of the ensemble GNN model was tested rigorously using the method of five-fold cross-validation. The testing protocol included a comprehensive array of metrics, such as accuracy, F1-score, precision, recall, and loss, to achieve a significant and multifaceted prediction accuracy of the ensemble model. Model uncertainty analysis and error analysis, both were performed to get further understanding about the approach’s strengths and weaknesses.

5.3.2.1 Cross-Validation Performance

Throughout all five rounds of cross-validation, our ensemble model maintained a high level of efficacy. We noted that the performance of the model was consistent for all five folds. In addition, the cross-validation-achieved scores indicated a high level of stability for the model’s predictive accuracy, ensuring minimal variation

achieved across the performed data partitions. Consequently, this indicates a strong ability for effective generalization to new samples, an important consideration within this field. We came across that the average cross-validation accuracy for all five folds was 93.41%, with individual fold performance varying from 89.94% to 95.52%. The limited variation in performance serves as a sign of the model’s stability against changes in data splits, indicating a low likelihood of overfitting to particular data subsets. Additionally, our average validation F1 score was 92.86%, which further illustrates the model’s balanced predictive performance, maintaining a suitable equilibrium between precision and recall, regardless of the inherent disparity between classes in the dataset. Finally, out of all the folds, we obtained the highest validation accuracy of 95.52% and the highest validation F1 score of 94.12% in our ensemble model. Training and validation loss for all the folds have a consistent decreasing trend, indicating optimization and model convergence. Difference between training and validation loss is minimal, which means that there isn’t significant overfitting despite the small dataset size. The standard deviation between-folds is low, indicating strong performance and low sensitivity to splitting of the data.

A thorough analysis of the cross-validation results is shown down below:

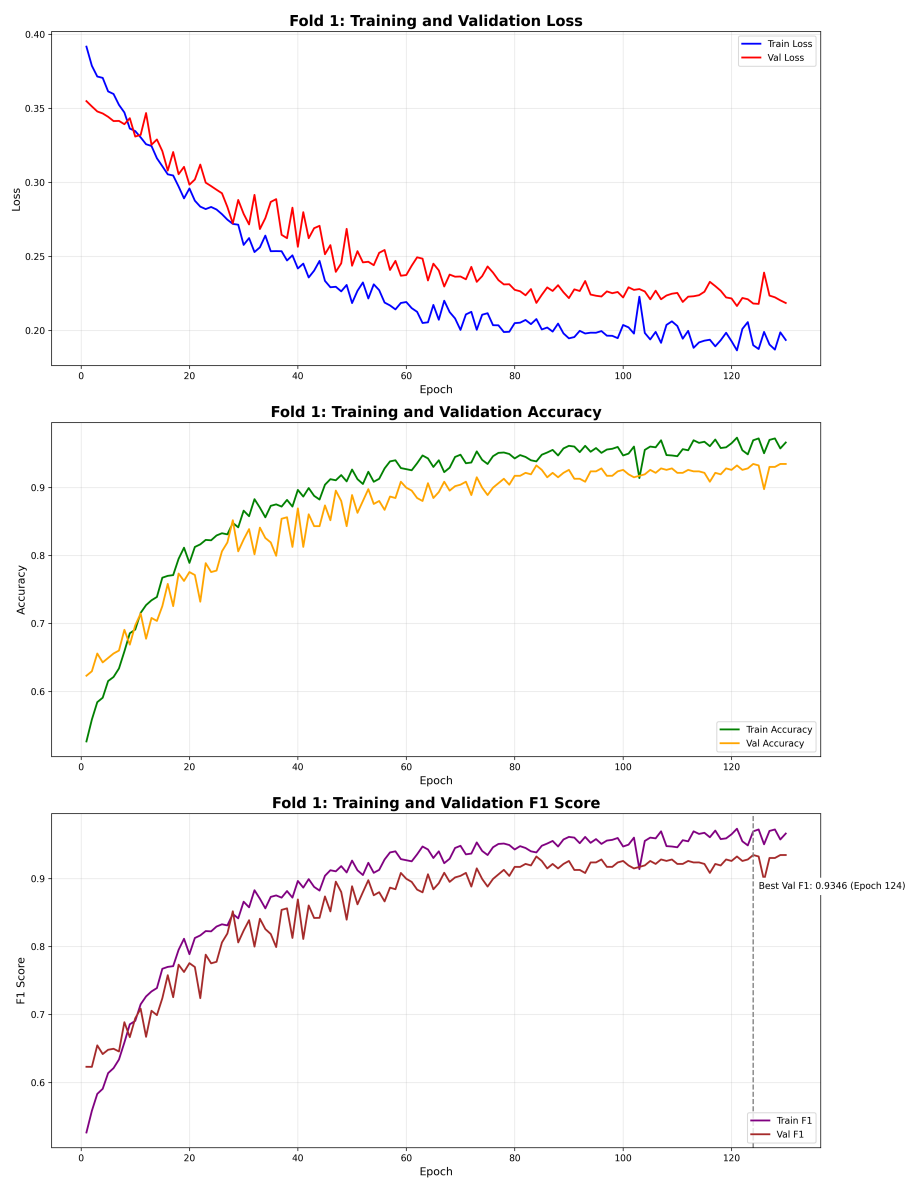


Figure 5.2: Graph Representation of Fold-1

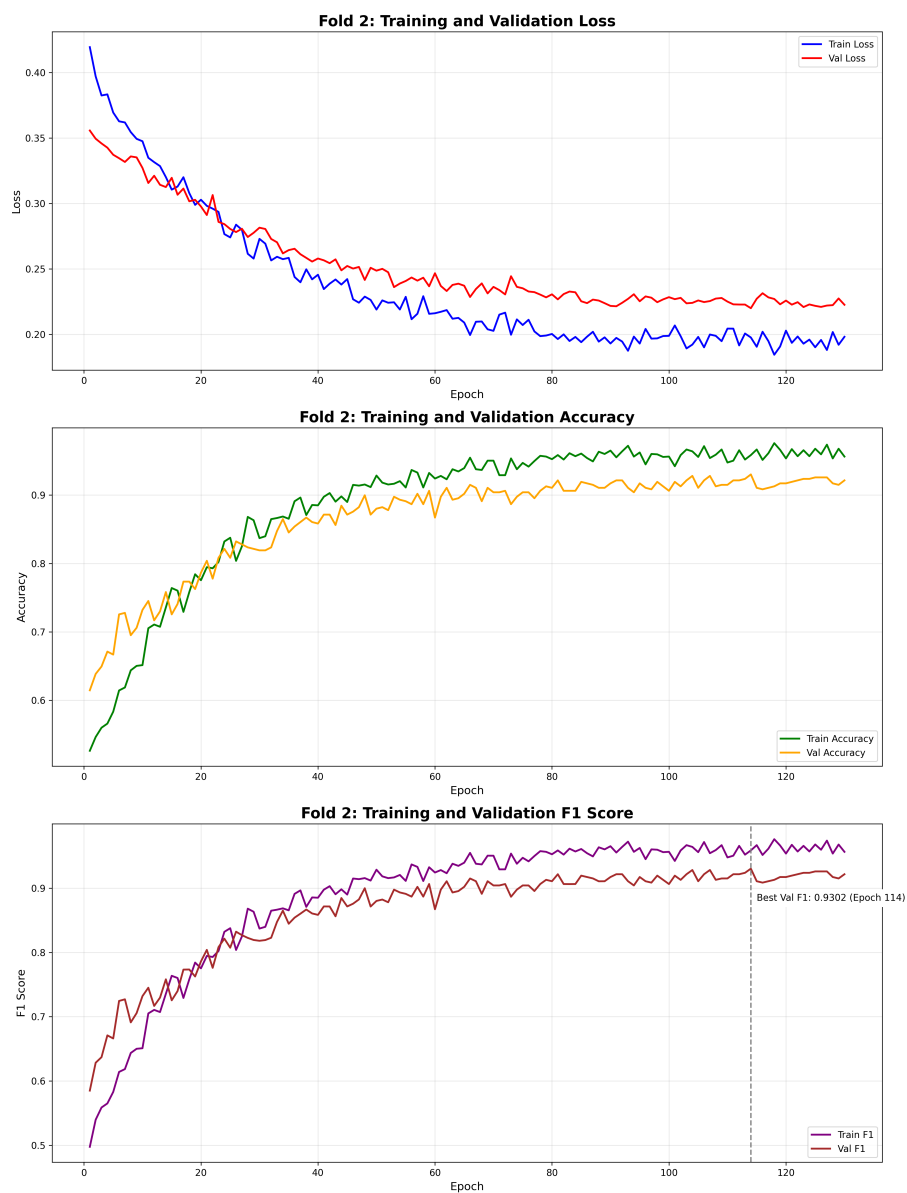


Figure 5.3: Graph Representation of Fold-2

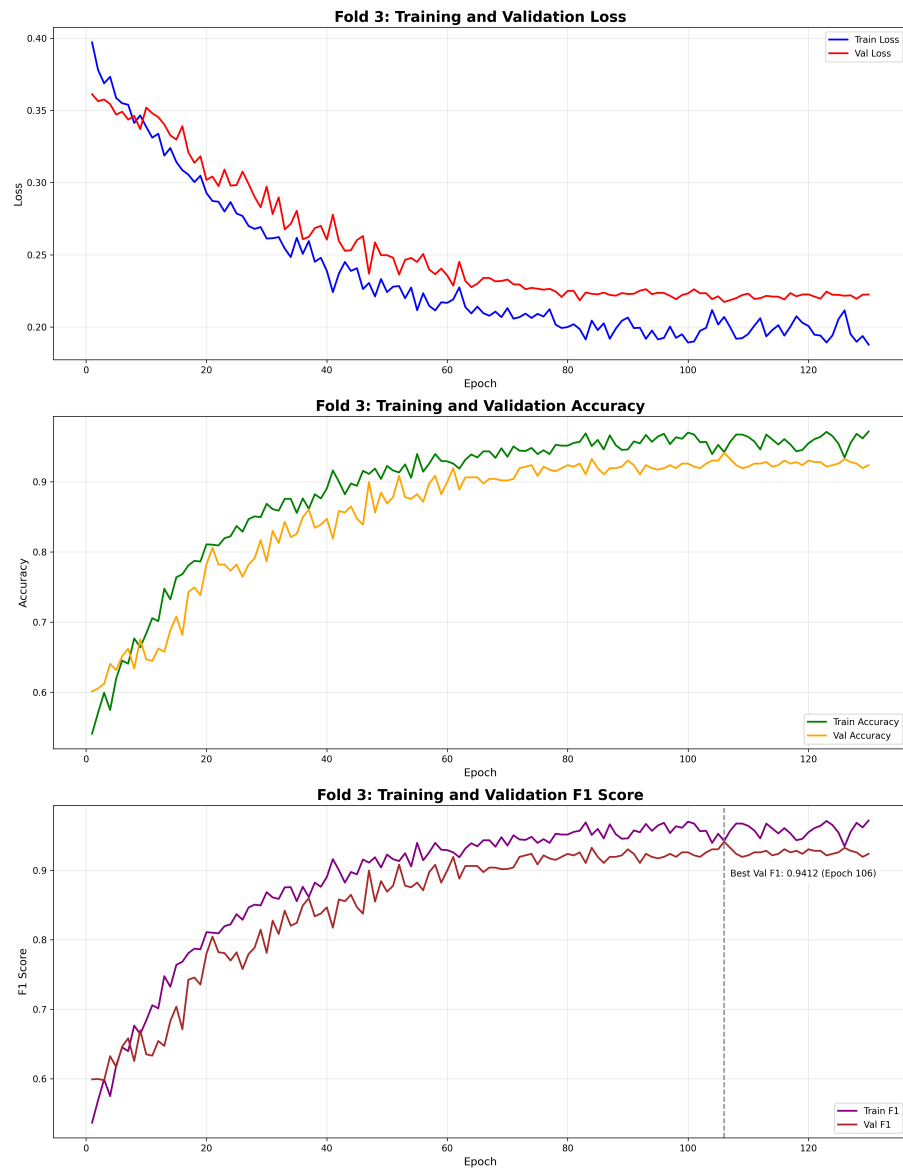


Figure 5.4: Graph Representation of Fold-3

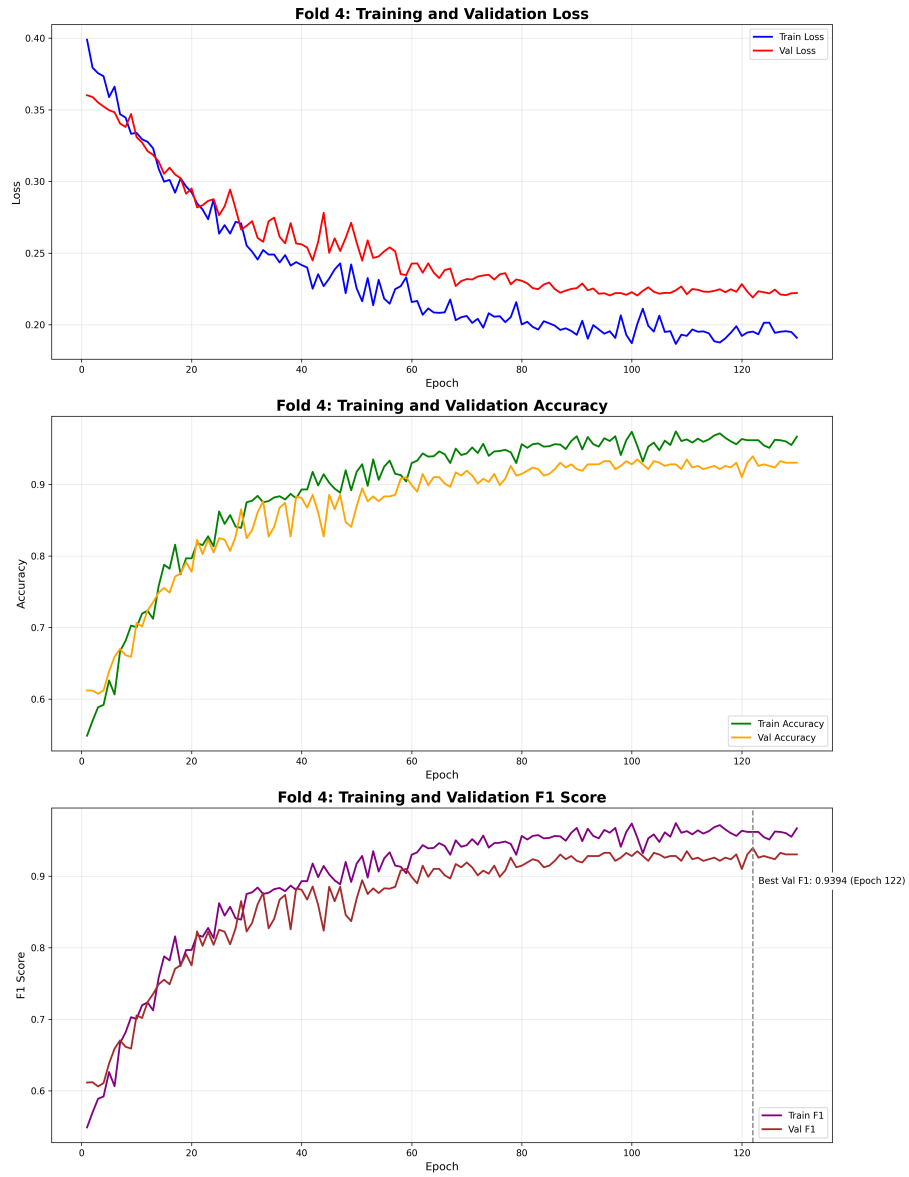


Figure 5.5: Graph Representation of Fold-4

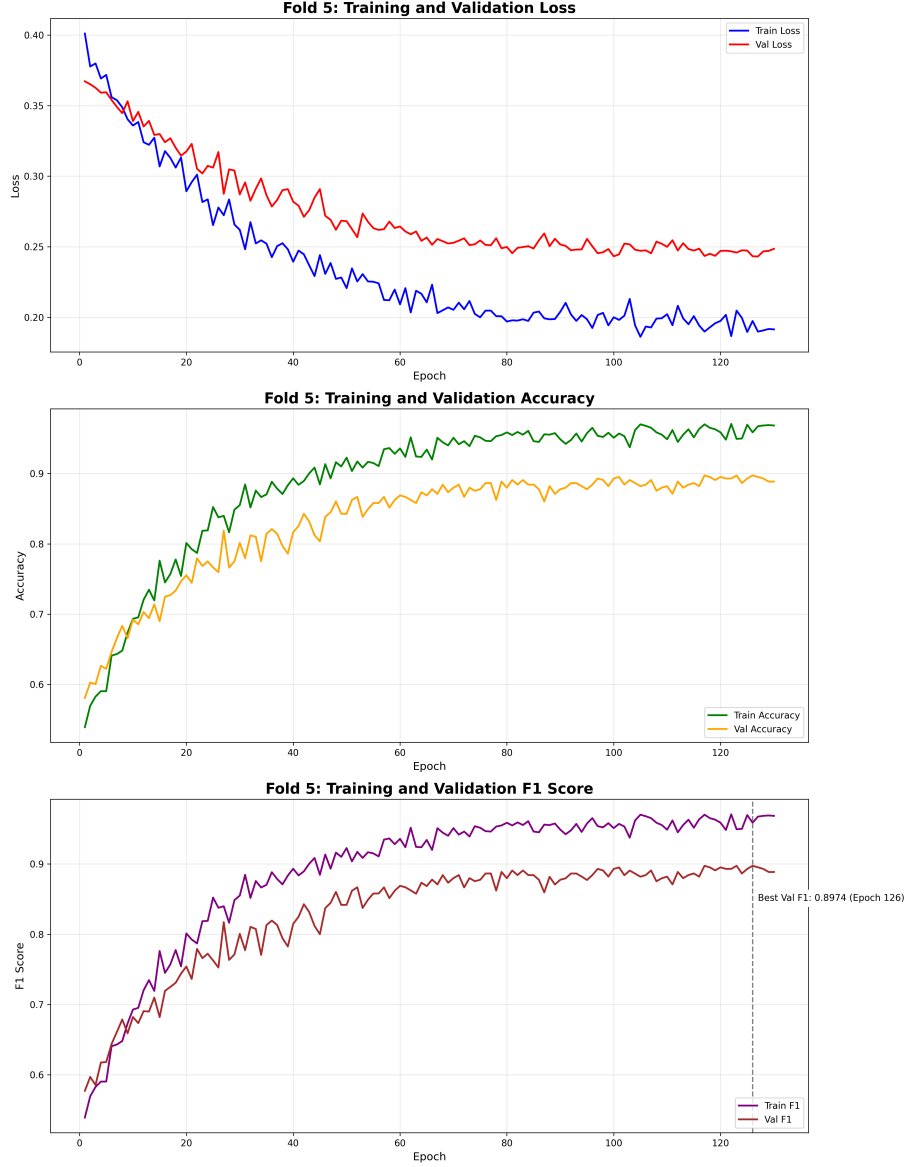


Figure 5.6: Graph Representation of Fold-5

Fold	Val Accuracy	Val F1-score	Val Precision	Val Recall	Val Loss	Val Uncertainty
1	0.9386	0.9346	0.9306	0.9307	0.2199	0.599
2	0.9372	0.9302	0.9271	0.9172	0.2398	0.553
3	0.9552	0.9412	0.9344	0.9324	0.2197	0.280
4	0.9399	0.9394	0.9265	0.9260	0.2216	0.262
5	0.8994	0.8974	0.8959	0.8836	0.2242	0.599
Mean	0.9341	0.9286	0.9229	0.9179	0.2251	0.459

Table 5.2: Cross-validation performance metrics for the ensemble GNN model

We observed that the model consistently performed similarly on all four evaluation metrics: recall, accuracy, F1 score, and precision. In terms of both memory and precision, it attained a well-rounded classification effectiveness with integrity. We see a demonstration of strong alignment between precision and recall, which ensures the fact that the model carefully works to minimize false positives while accurately identifying true cases.

An examination of the training process revealed distinct levels of learning. During the initial epochs, the model exhibited rapid improvement. During the initial 30 epochs, validation accuracy increased from approximately 60% to surpass 80%. Subsequent to the initial phase of rapid convergence, a period of more moderate growth ensued, characterized by modest performance improvements. Throughout the training procedure, the training metrics consistently outperformed the validation metrics. Ultimately, the training accuracy attained approximately 96–97% across all folds. The tight correlation between the training and validation metrics reflects the model’s reliability and its generalization capabilities. Upon analyzing the training curves for each fold, we noted that the learning patterns exhibited consistency across every scenario, which is favorable for the subsequent process. The convergence rates and final performance showed slight differences, as we see from the data. This inline outcome illustrates the reliability of the training methodology and the ensemble architecture across different data segments presented.

5.3.2.2 Fine-Tuning Performance

The fine-tuning procedure, applied after the primary training process, had varying levels of effectiveness across the five folds. In all of the five folds, fine-tuning led to statistically significant performance improvements in validation F1 score and validation accuracy. For example, in fold 4, the application of fine-tuning improved the validation F1 score from 93.18% to 93.94%. If any of the folds did not improve performance in fine tuning then the best models were obtained from the main training phase. The observed pattern of selective improvement indicates that the effectiveness of fine-tuning is contingent upon the distinct attributes of the validation set and the convergence state achieved during the primary training phase. Despite this variation, the conservative fine-tuning approach, which was marked by reduced learning rates ($2e-5$) and a stricter application of gradient clipping (maximum norm at 1.0), effectively suppressed performance degradation, hence ensuring that the fine-tuning process was, at least, performance-neutral. This safeguarding function proved its worth in maintaining the integrity of the high-performing models that resulted from the main training stage.

5.3.2.3 Uncertainty-Threshold Performance

As summarized in Table 5.2, predictive uncertainty on the mean was moderate across all the folds, and lower uncertainty generally maps to higher accuracy and F1-score. This type of behavior implies that the model is well-calibrated and able to distinguish between confident and ambiguous cases. Essentially, To better understand the pragmatic usefulness of uncertainty estimation, we compared performance across uncertainty thresholds (table 5.3). By considering only those predictions made with uncertainty under a certain threshold, the model had perfect accuracy, F1-score, precision, and recall (all 1.00) at the lowest thresholds. But We noticed it had extremely low coverage. As the threshold increased, the coverage improved, but the performance measures declined gradually. For predictions with uncertainty below 0.3 (very high confidence), the model achieved perfect performance: Accuracy 1.00, F1-score 1.00, Precision 1.00, Recall 1.00, though this only covered 2% of the data. At a more lenient uncertainty threshold of 0.6, the model maintained strong performance: Accuracy 0.70, F1-score 0.70, Precision 0.70, Recall 0.70, with much broader

coverage of 83%. This trade-off illustrates the model’s ability of identifying very confident predictions, which can effectively be given priority for automatic decision-making, whereas less certain cases are reserved for expert doctor review. This mode of operation is of particular usefulness in clinical pathways, where minimizing risk is of the highest priority. From a practical viewpoint, extremely uncertain predictions may be flagged for further review by an expert or additional diagnostic tests to avoid misclassification for high-risk clinical cases. But overall, the integration of uncertainty estimation not only enhances the model’s reliability and interpretability but also enables risk-aware deployment in real-world clinical practice.

Threshold	Accuracy	F1-score	Precision	Recall	Coverage
0.2	1.0000	1.0000	1.0000	1.0000	0.00
0.3	1.0000	1.0000	1.0000	1.0000	0.02
0.4	0.7568	0.6237	0.6237	0.6237	0.12
0.5	0.7208	0.7068	0.7071	0.7064	0.47
0.6	0.6982	0.6980	0.6986	0.6993	0.83

Table 5.3: Uncertainty-threshold performance metrics and coverage.

5.4 Confusion Matrix and Error Analysis

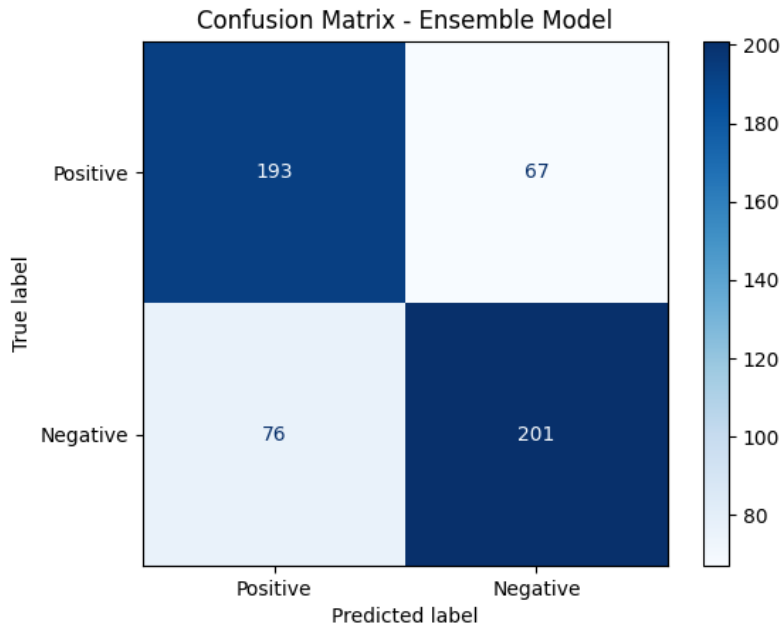


Figure 5.7: Confusion matrix for the ensemble model

Figure 5.7 presents the confusion matrix of the ensemble model, averaged across all validation predictions. The model is nicely balanced in terms of sensitivity and specificity, with 193 true positives, 201 true negatives, 67 false negatives, and 76 false positives. Moreover, we also notice that there are few misclassifications for both classes which comes to show that the model is not biased towards positive or

negative examples in the dataset. Both the false negative and false positive ratios are low as we noticed, and this is beneficial in the clinical environment because they help to avoid missed diagnoses and inappropriate interventions.

5.5 Key Findings and Significance

Our ensemble GNN model shows reliable performance as it is supported by high and stable cross-validation metrics, steady learning curves, and consistent error rates. The model’s ability to maintain good performance in all the folds, together with its low misclassification values, highlights its potential for reliable integration into clinical decision support systems. Furthermore, uncertainty estimation added an additional level of insight into the model’s confidence in predictions. Mean uncertainty values were modest for training and validation, showing the ensemble could recognize samples where predictions were less confident. This is especially useful in a clinical setting, where model uncertainty can be utilized to direct decision support and flag cases for additional expert review. The training curves, as drawn by the plotting utilities, show systematic decrease of loss and corresponding improvement in performance measures, with minimal signs of overfitting. Early stopping and fine-tuning methods also assisted in model stability and model checkpoint selection in an optimum way. Upon comparison we find that, the ensemble method combined with multiple graph neural network architectures performed better than individual models in both accuracy and stability. Our ensemble model achieved a highest validation F1 score of 94.12%, while the individual GraphMAE2 model displayed a validation F1 score of 69%. Despite the overall strong performance, there was some variation between the different folds in the ensemble model, but this may have resulted from the small sample and residual class imbalance. Finally, the ability of the ensemble model to maintain high validation scores across all the folds despite this demonstrates its potential for reliable colposcopy image classification.

Chapter 6

Limitations And Future Work

6.1 Limitations

Despite the comprehensive design and implementation of our colposcopy image analysis pipeline, several significant limitations existed during the process of this research. The foremost challenge among all the problems was the lack of availability of high-quality, publicly available colposcopy datasets. Efforts to acquire additional data through reaching out to the domain experts and scientists resulted in little progress, with the derived dataset being small in size. Not only did this unavailability restrict the diversity of the data but also cause an extremely imbalanced distribution of the classes, especially for the original histopathology-based categorization, with some classes containing only one or only two instances. Thus, we were compelled to drop the finer detail of the histopathology classification in favor of a binary HPV status classification that although was better balanced, but reduced the clinical granularity and the potential usefulness of the model.

Feature extraction from colposcopy images posed another substantial limitation. Standard deep learning architectures such as ResNet50 and DeepLabV3+, which have demonstrated strong performance on natural images and even other medical imaging modalities, failed to capture the subtle and domain-specific features necessary for effective colposcopy analysis [51]. So, we made a custom feature extractor, which mainly includes 5 clinical categories, for example, acetowhite features, vascular features, tissue features, etc. Because of that, we could successfully extract the important information from the images. Furthermore, in feature extraction, we were having a problem extracting information from the cervix zone. Because if the model gets information from outside the cervix zone, it would be detrimental to detect whether that is cervical cancer positive or negative. So, we manually annotated the cervix zone and used that JSON in our pre-processing phase to correctly detect and extract features from the cervix zone.

The existing study in this domain is limited, with most studies relying on traditional image-based or tabular approaches, and only a handful exploring the use of graph neural networks (GNNs) for medical imaging tasks. Not having enough established benchmarks meant that our research had to include extensive experimentation and iterative trial and error so that the most effective models could be selected for our ensemble.

6.2 Future Work

Our ensemble GNN model exhibits encouraging performance on cervical cancer diagnosis through a variety of graph architectures. Our proposed work involves cervix-aware attention and uncertainty estimation for enhanced model interpretability. Moreover, our future research focuses on a variety of domains for enhanced clinical utility and diagnosis robustness.

- We'll try to develop advanced explainability tools for GNN-based image interpretation. Specifically, we intend to visually demonstrate attention distributions along graph structures, generate node- and region-level saliency maps, as well as quantify the impact of particular subgraph elements. We envision these methods to allow for actionable decision support for clinicians, promote clinical trust, and support seamless incorporation of AI-augmented diagnostics into routine workflows.
- The existing pipeline depends on manual annotations of the cervix area to direct feature weighting and attention. We also intend to incorporate auto-segmentation of the cervix into our pipeline, utilizing models such as U-Net or transformer-based architectures. This will eliminate manual annotations, permit end-to-end automation, and improve the scalability and generalizability of our framework to other clinical datasets and applications.
- To simulate disease progression for patients with more than one visit for colposcopic diagnosis, we intend to incorporate temporal modeling techniques, such as temporal graph neural networks or recurrent architectures. These enhancements will allow us to view temporal dynamics on cervical tissue development and facilitate more early and accurate intervention strategies.
- Usage of Skippool can reduce the number of nodes which can capture multiple node features. Even one node can capture all nodes features, which can ensure lower computational power [53]. So we will try to explore in that direction as well.

In the future, our ongoing efforts will continue to improve predictive precision and clinical validity and enable early diagnosis and tailored treatment of cervical cancer.

Chapter 7

Conclusion

We introduced a tailored ensemble framework in our study designed specifically for the detection of cervical cancer, combining the unique capabilities of various graph-based neural networks. In recent times, Graph neural networks offer diagnostic quality and are a viable option for medical image segmentation. By integrating GraphMAE2, GAT, GIN, and GCN models with new improvements, we were able to solve persistent challenges such as medical data imbalance and uncertainty of predictions inherent to medical image classification. The ensemble strategy in our research utilizes advanced techniques such as focal loss and label smoothing for balancing skewed class distributions and uncertainty estimation, coupled with temperature scaling for enhanced model calibration. Our carefully cross-validated and hyperparameter-tuned ensemble also improved the performance of the ensemble to an accuracy of 95.52% and F1 score of 94.12% on validation. In comparison to standard single-model approaches, our solution outperformed exemplary markers of accuracy, F1 score, and confidence predictions consistently. Moreover, the ensemble’s ability to make calibrated confidence predictions provides interpretability and reliability, which is critical for clinical deployment. Within this research, we identify graph ensemble models for their potential to advance the boundaries of medical image processing. In summary, our proposed ensemble model provides a compelling solution for clinical contexts, for high-stakes diagnosis tasks by offering both reliability and performance.

Bibliography

- [1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14124313>.
- [2] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, *Patch-based convolutional neural network for whole slide tissue image classification*, 2016. arXiv: 1504.07947 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1504.07947>.
- [3] T. N. Kipf, *Graph convolutional networks*, <https://tkipf.github.io/graph-convolutional-networks/>, Blog post; overview and introduction to GCNs, 2016. (visited on 06/18/2025).
- [4] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, 2017. arXiv: 1609.02907 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1609.02907>.
- [5] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, “Deep learning for image-based cancer detection and diagnosis — a survey,” *Pattern Recognition*, vol. 83, May 2018. DOI: 10.1016/j.patcog.2018.05.014.
- [6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, *Graph attention networks*, 2018. arXiv: 1710.10903 [stat.ML]. [Online]. Available: <https://arxiv.org/abs/1710.10903>.
- [7] B. Ghojogh, M. Samad, S. Mashhadi, *et al.*, *Feature selection and feature extraction in pattern analysis: A literature review*, May 2019. DOI: 10.48550/arXiv.1905.02845.
- [8] X. Liu, L. Faes, A. Kale, *et al.*, “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis,” *The Lancet Digital Health*, vol. 1, Sep. 2019. DOI: 10.1016/S2589-7500(19)30123-2.
- [9] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, *How powerful are graph neural networks?* 2019. arXiv: 1810.00826 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1810.00826>.
- [10] M. Adnan, S. Kalra, and H. R. Tizhoosh, *Representation learning of histopathology images using graph neural networks*, 2020. arXiv: 2004.07399 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2004.07399>.
- [11] S. Kaur Saini, V. Bansal, R. Kaur, and M. Juneja, “Colponet for automated cervical cancer screening using colposcopy images,” *Machine Vision and Applications*, vol. 31, Mar. 2020. DOI: 10.1007/s00138-020-01063-8.

- [12] Y. Ou, Y. Xue, Y. Yuan, *et al.*, *Semi-supervised cervical dysplasia classification with learnable graph convolutional network*, 2020. arXiv: 2004.00191 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2004.00191>.
- [13] C. Rajarao and R. Singh, “Improved normalized graph cut with generalized data for enhanced segmentation in cervical cancer detection,” *Evolutionary Intelligence*, vol. 13, pp. 3–8, 2020. DOI: 10.1007/s12065-019-00226-5. [Online]. Available: <https://doi.org/10.1007/s12065-019-00226-5>.
- [14] P. Xue, M. Ng, and Y. Qiao, “The challenges of colposcopy for cervical cancer screening in lmics and solutions by artificial intelligence,” *BMC Medicine*, vol. 18, Dec. 2020. DOI: 10.1186/s12916-020-01613-x.
- [15] C. Yuan, Y. Yao, B. Cheng, *et al.*, “The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images,” *Scientific Reports*, vol. 10, Jul. 2020. DOI: 10.1038/s41598-020-68252-3.
- [16] J. Zhou, G. Cui, S. Hu, *et al.*, “Graph neural networks: A review of methods and applications,” *AI Open*, vol. 1, pp. 57–81, 2020, ISSN: 2666-6510. DOI: <https://doi.org/10.1016/j.aiopen.2021.01.001>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666651021000012>.
- [17] W. Li, “Analysis of object detection performance based on faster r-cnn,” *Journal of Physics: Conference Series*, vol. 1827, p. 012085, Mar. 2021. DOI: 10.1088/1742-6596/1827/1/012085.
- [18] S. Lin *et al.*, “Worldwide trends in cervical cancer incidence and mortality, with predictions for the next 15 years,” *Cancer*, vol. 127, pp. 4030–4039, 2021. DOI: 10.1002/cncr.33652.
- [19] L. Qiu, H. Li, M. Wang, and X. Wang, “Gated graph attention network for cancer prediction,” *Sensors*, vol. 21, no. 6, p. 1938, 2021. DOI: 10.3390/s21061938. [Online]. Available: <https://doi.org/10.3390/s21061938>.
- [20] J. Shi, R. Wang, Y. Zheng, Z. Jiang, H. Zhang, and L. Yu, “Cervical cell classification with graph convolutional network,” *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105807, 2021, Epub 2020 Oct 22. DOI: 10.1016/j.cmpb.2020.105807.
- [21] J. Shi, R. Wang, Y. Zheng, Z. Jiang, H. Zhang, and L. Yu, “Cervical cell classification with graph convolutional network,” *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105807, Jan. 2021. DOI: 10.1016/j.cmpb.2020.105807.
- [22] S. Shintaro, K. Yoshii, T. Hara, *et al.*, “Multi-task deep learning model for classification of dental implant brand and treatment stage using dental panoramic radiograph images,” *Biomolecules*, vol. 11, p. 815, May 2021. DOI: 10.3390/biom11060815.
- [23] O. Tapera *et al.*, “Gaps and opportunities for cervical cancer prevention, diagnosis, treatment and care: Evidence from midterm review of the zimbabwe cervical cancer prevention and control strategy (2016–2020),” *BMC Public Health*, vol. 21, pp. 1–13, 2021. DOI: 10.1186/s12889-021-10767-3.

- [24] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang, “Graph neural networks and their current applications in bioinformatics,” *Frontiers in Genetics*, vol. 12, p. 690 049, Jul. 2021. DOI: 10.3389/fgene.2021.690049.
- [25] M. Bilal, R. Jewsbury, R. Wang, *et al.*, *An aggregation of aggregation methods in computational pathology*, 2022. arXiv: 2211.01256 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2211.01256>.
- [26] Y. Cao, H. Ma, Y. Fan, *et al.*, “A deep learning-based method for cervical transformation zone classification in colposcopy images,” *Technology and Health Care*, vol. 31, pp. 1–12, Sep. 2022. DOI: 10.3233/THC-220141.
- [27] L. Habtemariam, E. Taye, and G. Simegn, “Cervix type and cervical cancer classification system using deep learning techniques,” *Medical Devices: Evidence and Research*, vol. 15, pp. 163–176, Jun. 2022. DOI: 10.2147/MDER.S366303.
- [28] Z. Hou, X. Liu, D. Yuxiao, H. Yang, C. Wang, and J. Tang, *Graphmae: Self-supervised masked graph autoencoders*, May 2022. DOI: 10.48550/arXiv.2205.10803.
- [29] W. Huang, S. Sun, Z. Yu, S. Lu, and H. Feng, “Chronic cervicitis and cervical cancer detection based on deep learning of colposcopy images toward translational pharmacology,” *Frontiers in Pharmacology*, vol. 13, 2022, ISSN: 1663-9812. DOI: 10.3389/fphar.2022.911962.
- [30] T. Huynh, A. Nibali, and Z. He, “Semi-supervised learning for medical image classification using imbalanced training data,” *Computer Methods and Programs in Biomedicine*, vol. 216, p. 106 628, 2022, ISSN: 0169-2607. DOI: <https://doi.org/10.1016/j.cmpb.2022.106628>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016926072200013X>.
- [31] R. Pramanik, M. Biswas, S. Sen, L. Souza Jr, J. Papa, and R. Sarkar, “A fuzzy distance-based ensemble of deep models for cervical cancer detection,” *Computer Methods and Programs in Biomedicine*, vol. 219, p. 106 776, Jun. 2022. DOI: 10.1016/j.cmpb.2022.106776.
- [32] M. A. Devi, R. Ezhilarasie, K. S. Joseph, K. Kotecha, A. Abraham, and S. Vairavasundaram, “An improved boykov’s graph cut-based segmentation technique for the efficient detection of cervical cancer,” *IEEE Access*, vol. 11, pp. 77 636–77 647, 2023. DOI: 10.1109/ACCESS.2023.3295833.
- [33] Z. Hou, Y. He, Y. Cen, *et al.*, *Graphmae2: A decoding-enhanced masked self-supervised graph learner*, Apr. 2023. DOI: 10.48550/arXiv.2304.04779.
- [34] M. Kalbhor, S. Shinde, S. Lahade, and T. Choudhury, “Deepcervicancer - deep learning-based cervical image classification using colposcopy and cytology images,” *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 9, Oct. 2023. DOI: 10.4108/eetpht.9.3473.
- [35] M. Kalbhor and V. Swati, “Comparative analysis of deep learning pre-trained models and machine learning classifiers for accurate cervical cancer diagnosis using colposcopy images,” *BMC Medicine*, Dec. 2023.
- [36] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, “Medical image analysis using deep learning algorithms,” *Frontiers in Public Health*, vol. 11, p. 1 273 253, Nov. 2023. DOI: 10.3389/fpubh.2023.1273253.

- [37] J. M B and P. Subashini, "Improvising grading of cervical cancer using quality assessment method in smart colposcopy images," *Measurement: Sensors*, vol. 27, p. 100788, May 2023. DOI: 10.1016/j.measen.2023.100788.
- [38] W. A. Mustafa, S. Ismail, F. S. Mokhtar, H. Alquran, and Y. Al-Issa, "Cervical cancer detection techniques: A chronological review," *Diagnostics (Basel)*, vol. 13, no. 10, p. 1763, May 2023. DOI: 10.3390/diagnostics13101763.
- [39] E. Pichka. "Graph attention networks paper explained with illustration and pytorch implementation." Web blog post. (2023), [Online]. Available: <https://epichka.com/blog/2023/gat-paper-explained/> (visited on 06/18/2025).
- [40] N. Youneszade, M. Marjani, and S. Ray, "A predictive model to detect cervical diseases using convolutional neural network algorithms and digital colposcopy images," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2023. DOI: 10.1109/ACCESS.2023.3285409.
- [41] N. Zaki, W. Qin, and A. Baby, "Graph-based methods for cervical cancer segmentation: Advancements, limitations, and future directions," *AI Open*, vol. 4, Aug. 2023. DOI: 10.1016/j.aiopen.2023.08.006.
- [42] H. Ahmadzadeh Sarhangi, D. Beigifard, E. Farmani, and H. Bolhasani, "Deep learning techniques for cervical cancer diagnosis based on pathology and colposcopy images," *Informatics in Medicine Unlocked*, vol. 47, p. 101503, 2024, ISSN: 2352-9148. DOI: <https://doi.org/10.1016/j.imu.2024.101503>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914824000595>.
- [43] S. Kumar M, D. Francis, S. Srinivasan, V. Khatavkar, K. P, and M. Shah, "Enhancing cervical cancer detection and robust classification through a fusion of deep learning models," *Scientific Reports*, vol. 14, May 2024. DOI: 10.1038/s41598-024-61063-w.
- [44] B. Omodunbi, A. Soladoye, A. Esan, S. Emesowum, T. Olowo, and O. Ojelabi, "Detection of cervical cancer using deep transfer learning," *Dutse Journal of Pure and Applied Sciences*, vol. 10, pp. 29–37, Mar. 2024. DOI: 10.4314/dujopas.v10i1a.4.
- [45] M. Qathrady, A. Shaf, T. Ali, *et al.*, "A novel web framework for cervical cancer detection system: A machine learning breakthrough," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2024. DOI: 10.1109/ACCESS.2024.3377124.
- [46] G. Su, H. Wang, Y. Zhang, W. Zhang, and X. Lin, "Simple and deep graph attention networks," *Knowledge-Based Systems*, vol. 293, p. 111649, 2024, ISSN: 0950-7051. DOI: <https://doi.org/10.1016/j.knosys.2024.111649>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705124002843>.
- [47] A. G. Vrahatis, K. Lazaros, and S. Kotsiantis, "Graph attention networks: A comprehensive review of methods and applications," *Future Internet*, vol. 16, no. 9, p. 318, 2024. DOI: 10.3390/fi16090318. [Online]. Available: <https://doi.org/10.3390/fi16090318>.
- [48] N. Youneszade, M. Marjani, and D. Shafiq, "Exploring the impact of increasing the number of classes on the performance of cervical cancer detection models using deep learning and colposcopy," *Journal of Engineering Science and Technology*, vol. 19, pp. 629–647, Apr. 2024.

- [49] D. Buterez, J. P. Janet, D. Oglic, and P. Liò, “An end-to-end attention-based approach for learning on graphs,” *Nature Communications*, vol. 16, no. 1, p. 5244, 2025. DOI: 10.1038/s41467-025-60252-z. [Online]. Available: <https://doi.org/10.1038/s41467-025-60252-z>.
- [50] P. Chatterjee, S. Siddiqui, R. S. A. Kareem, and S. R. Rao, “Multi-modal graph neural networks for colposcopy data classification and visualization,” *Cancers*, vol. 17, no. 9, p. 1521, 2025. DOI: 10.3390/cancers17091521. [Online]. Available: <https://doi.org/10.3390/cancers17091521>.
- [51] P. Chatterjee, S. Siddiqui, R. S. A. Kareem, and S. Rao, “Attention-enhanced lightweight architecture with hybrid loss for colposcopic image segmentation,” *Cancers*, vol. 17, no. 5, 2025, ISSN: 2072-6694. DOI: 10.3390/cancers17050781. [Online]. Available: <https://www.mdpi.com/2072-6694/17/5/781>.
- [52] S. Hu, N. Bhasker, G. Montufar, J. N. Kather, L. Seifert, and S. Speidel, “Pg-gat: A complete graph model for cancer detection and subtyping in whole slide images analysis,” in *International Conference on Learning Representations (ICLR)*, Submitted to ICLR 2025, last modified: 05 Feb 2025, 2025. [Online]. Available: <https://openreview.net/forum?id=MOCEoNsjEx>.
- [53] S. Imaduwege, “Skipool: Improved sparse hierarchical graph pooling with differentiable exploration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 17 546–17 554. DOI: 10.1609/aaai.v39i17.33929. [Online]. Available: <https://doi.org/10.1609/aaai.v39i17.33929>.
- [54] H. Kaur, R. Sharma, and J. Kaur, “Comparison of deep transfer learning models for classification of cervical cancer from pap smear images,” *Scientific Reports*, vol. 15, p. 3945, 2025. DOI: 10.1038/s41598-024-74531-0.
- [55] A. Sadeghi, F. Hajati, A. Argha, N. H. Lovell, M. Yang, and H. Alinejad-Rokny, *Interpretable graph-based models on multimodal biomedical data integration: A technical review and benchmarking*, 2025. arXiv: 2505.01696 [q-bio.GN]. [Online]. Available: <https://arxiv.org/abs/2505.01696>.