DEEP LEARNING ENABLED FRAMEWORK FOR EXPLAINABLE COLORECTAL POLYP CLASSIFICATION

FARAH YOUNAS

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

2025

School of Engineering, Computer and Mathematical Sciences

Abstract

Cancer remains a leading global cause of death, with its burden increasing worldwide. Early detection and classification of polyps are crucial for reducing cancer risk and improving outcomes. Accurate classification of colorectal cancer (CRC) is vital for effective healthcare and diagnosis. Automated medical imaging aids in these diagnoses, but pathologists often face challenges like inexperience and visual fatigue. This research aims to develop a sensitive virtual biopsy tool to assist pathologists in overcoming issues related to the quality, diversity, and variability of biomedical data.

This thesis aims to develop and evaluate a novel framework for the efficient and accurate classification of CRC polyps from endoscopy images. It involves studying the existing techniques, designing and implementing the classification framework, and validating the model to enhance sensitivity, accuracy, and clinical applicability. This research utilises deep learning models, including CNNs, CGANs, Transformer-based models, XAI models, and Nonlinear Dimensionality Reduction. It incorporates transfer and ensemble learning, synthetic image generation, fine-grained feature extraction, model explainability, and manifold learning with regularisation to enhance classification performance. The approach also includes a comprehensive pipeline for data preprocessing, model training, and performance evaluation on real-world datasets.

This novel deep learning framework achieves state-of-the-art results in CRC polyp classification by generating high-quality synthetic images and extracting detailed features for an explainable model. It uses nonlinear dimensionality reduction to handle high dimensionality and enhance performance. The model demonstrates an average accuracy improvement of 18% and a 17% increase in sensitivity compared to traditional methods, with sensitivities of 0.95, 0.96, and 0.96 on the UCI, PICCOLO, and CPDC datasets, respectively. Its strong performance across various metrics suggests it is a promising tool for clinical integration, aiding gastroenterologists in polyp classification for polypectomy.

This thesis significantly advances CRC image classification by developing a framework that addresses multiple data challenges and improves classification performance. The model shows potential for clinical use as a virtual biopsy tool to aid pathologists and enhance patient outcomes. Future work will focus on integrating real-time data detection, handling inaccurate clinical data, and developing a semi-supervised learning model to train with limited labelled data.

Contents

Abstract **2**

Attestation of Authorship **11**

Acknowledgements **12**

List of Abbreviations **13**

1 Introduction **15**
1.1 Background 16
1.1.1 Cancer Detection and Classification 17
1.1.2 Deep Learning in Cancer Classification 21
1.2 Research Problems 24
1.2.1 Data Accessibility 24
1.2.2 Feature Extraction and Model Transparency 27
1.2.3 Nonlinear Dimensionality Reduction and Model Regularisation 34
1.2.4 Model Evaluation 36
1.3 Research Questions 38
1.4 Research Aims and Contributions 39
1.5 Research Methodology 41
1.6 Thesis Structure 44
1.7 Publications from Thesis 50

2 Literature Review **52**
2.1 Colorectal Cancer Classification 53
2.2 Synthetic Image Generation 73
2.3 Feature Extraction 80
2.4 Model Explainability 87
2.5 Feature Selection 92
2.6 Model regularisation and Image Classification 98
2.7 Problem Formulation 103
2.8 Summary of Chapter 113

3 Explainable Deep Learning Framework for CRC Polyp Classification 119
3.1 Methodological Framework . . . 120
3.1.1 Ensemble Learning Block (ELB) . . . 120
3.1.2 Synthetic Image Generation Block (SIGB) . . . 128
3.1.3 Deep Feature Extraction Block (DFEB) . . . 135
3.1.4 Model Explainable Block (MEB) . . . 144
3.1.5 Manifold Regularization Block (MRB) . . . 145
3.1.6 Concluding DNN Classifier . . . 148
3.2 Summary of Chapter . . . 153

4 Experiments on Real-world CRC Datasets 157
4.1 Datasets for Learning Model . . . 158
4.1.1 UCI Dataset . . . 158
4.1.2 PICCOLO Dataset . . . 158
4.1.3 CPDC Dataset . . . 159
4.2 Experimental Setup . . . 159
4.2.1 Scalability Considerations for Clinical Deployment . . . 161
4.3 Network Architecture . . . 162
4.4 Evaluation Metrics . . . 167
4.4.1 Classification Performance Analysis . . . 167
4.4.2 Reliability Analysis . . . 169
4.4.3 Image Quality and Diversity Analysis . . . 170
4.4.4 Manifold Learning Analysis . . . 172
4.5 Summary of Chapter . . . 174

5 Results and Discussion 179
5.1 Ensemble Learning Block . . . 180
5.1.1 Transfer Learning-Based Polyp Classification . . . 180
5.1.2 Ensemble-Based Polyp Classification . . . 185
5.1.3 Reliability Analysis . . . 196
5.1.4 Precision-Recall Analysis . . . 199
5.2 Synthetic Image Generation and Deep Feature Extraction . . . 209
5.2.1 FID, IS, PSNR and SSIM Analysis . . . 213
5.2.2 Performance Evaluation . . . 214
5.2.3 Specificity and Sensitivity Analysis . . . 215
5.2.4 Type I and Type II Error Analysis . . . 217
5.3 Model Explainability and Manifold regularisation . . . 218
5.4 Concluding Classifier . . . 226
5.5 Qualitative Analysis: Model Decision Making in Challenging Cases . . . 231
5.6 Summary of Chapter . . . 232

6 Conclusion **237**
6.1 Summary of Key Findings 247
6.2 Implications of Research 249
6.3 Limitations and Future Work 249
6.4 Road Map for Future Work 252

References **254**

List of Tables

Table 2.1: Review of detection and classification techniques of CRC diagnosis 74
Table 2.2: Review of cGAN, FEx, XAI and Manifold Learning in medical diagnosis 104

Table 3.1: Concordance determination of generated images 134

Table 4.1: Datasets details 162

Table 5.1: Experimental settings 181
Table 5.2: Learning rate: 0.001 and 100 epochs 183
Table 5.3: Results with ADAM optimiser, 0.001 and 0.005 learning rate, and 100 epochs 183
Table 5.4: Number of polyps per category 184
Table 5.5: Results with 0.001 learning rate and 100 epochs 186
Table 5.6: Results with SGDM optimiser and 100 epochs 187
Table 5.7: Optimized hyperparameter settings 187
Table 5.8: Performance evaluation of multiclass imbalanced and balanced dataset 188
Table 5.9: Performance of the base and ensemble classifier on imbalanced dataset 192
Table 5.10: Sensitivity and specificity of the base and ensemble classifier on imbalanced dataset 193
Table 5.11: Performance of the base and ensemble classifier on augmented dataset 193
Table 5.12: Sensitivity and specificity of the base and ensemble classifier on augmented dataset 194
Table 5.13: Performance evaluation of multiclass augmented dataset 194
Table 5.14: Performance of the base and ensemble classifier on binary class augmented dataset 195
Table 5.15: Performance evaluation of multiclass augmented dataset 203
Table 5.16: Performance of the base and ensemble classifier on binary class augmented dataset 204

Table 5.17: Sensitivity and specificity of the base and ensemble classifier on binary class augmented dataset 205
Table 5.18: Execution time comparison of datasets 205
Table 5.19: Comparative analysis of UCI and PICCOLO datasets with existing studies 207
Table 5.20: Comparative analysis of PICCOLO and CPDC datasets with existing studies 208
Table 5.21: Evaluation of concordant cDCGAN on UCI, PICCOLO, and CPDC datasets 209
Table 5.22: Confusion matrix of all datasets with and without concordant classifier integration 213
Table 5.23: Performance evaluation of updated UCI, PICCOLO and CPDC datasets with and without concordant classifier integration 214
Table 5.24: Performance evaluation of updated UCI, PICCOLO, and CPDC datasets with and without concordant classifier integration 215
Table 5.25: Performance comparison of proposed model 218
Table 5.26: Comparison between various manifold learning algorithms for the optimum components for UCI dataset 220
Table 5.27: Comparison between various manifold learning algorithms for the optimum components for PICCOLO dataset 223
Table 5.28: Comparison between various manifold learning algorithms for the optimum components for CPDC dataset 225
Table 5.29: Confusion matrix and performance evaluation on all validation datasets for proposed classification model 230
Table 5.30: Performance measures of proposed model for all datasets 230
Table 5.31: Advanced metric evaluation of proposed model for all datasets . . 231

List of Figures

Figure 3.1: Methodology of weighted-average ensemble learning block . . . 126
Figure 3.2: Framework of synthetic image generation block 133
Figure 3.3: Methodology of deep feature extraction block 143
Figure 3.4: Feature Extraction using Vision Transformer 143
Figure 3.5: Methodological framework for CRC classification 151

Figure 4.1: Sample polyp images of datasets 160

Figure 5.1: Accuracy and precision, optimiser: ADAM, LR: 0.001, epochs 50, 100 182
Figure 5.2: Recall and f1-score, optimiser: ADAM, LR: 0.001, 0.005, epochs: 50, 100 182
Figure 5.3: Accuracy and precision, optimisers: SGDM, LR: 0.001, 0.005, Epochs: 50, 100 185
Figure 5.4: Recall and f1-score, optimiser: SGDM, LR: 0.001, 0.005, Epochs: 50, 100 186
Figure 5.5: Weighted ensemble learning macro results comparison for imbalanced and balanced 189
Figure 5.6: Base and ensemble classifier performance on balanced dataset . . 191
Figure 5.7: Macro precision and false positive rate comparison of imbalanced and balanced UCI dataset 197
Figure 5.8: Macro precision and false positive rate comparison of imbalanced and balanced PICCOLO dataset 197
Figure 5.9: Reliability comparison using Cohen's kappa coefficient on UCI dataset 198
Figure 5.10: Reliability comparison using Cohen's kappa coefficient on PICCOLO dataset 198
Figure 5.11: AUC-ROC of PICCOLO Dataset 199
Figure 5.12: AUC-ROC of binary and multiclass classification on UCI dataset 199
Figure 5.13: Base and ensemble classifier performance on augmented dataset . 201
Figure 5.14: Precision-based comparison of imbalanced and augmented data . 201
Figure 5.15: False positive rate comparison of imbalanced and augmented data 202
Figure 5.16: CPDC dataset performance analysis 202

Figure 5.17: False positive and precision comparison of CPDC augmented data 203
Figure 5.18: Reliability comparison using Cohen's kappa coefficient 204
Figure 5.19: AUC-ROC of the proposed classifier on test data 206
Figure 5.20: MCC comparison of validation and test set 206
Figure 5.21: Network structure of the concordant cDCGAN generator 210
Figure 5.22: Network structure of the concordant cDCGAN discriminator . . . 211
Figure 5.23: Network structure of the concordant cDCGAN generator and discriminator combined 212
Figure 5.24: Concordant synthetic images of all datasets generated by the model 216
Figure 5.25: **(a)** Comparison of sensitivity, specificity, and **(b)** Type I, Type II error on UCI and PICCOLO dataset for non-concordant and concordant datasets 217
Figure 5.26: Generated explainable sample images for UCI dataset 219
Figure 5.27: Manifold learning visualizations on UCI dataset 221
Figure 5.28: Generated Explainable Sample Images for PICCOLO Dataset . . 222
Figure 5.29: Manifold learning visualizations on PICCOLO dataset 224
Figure 5.30: Generated explainable sample images for CPDC dataset 225
Figure 5.31: Manifold learning visualizations on CPDC dataset 227
Figure 5.32: ARI comparative visualization on UCI, PICCOLO, and CPDC dataset 228
Figure 5.33: NMI comparative visualization on UCI, PICCOLO, and CPDC dataset 228
Figure 5.34: FMI comparative visualization on UCI, PICCOLO, and CPDC dataset 228
Figure 5.35: Confusion matrix visualization of UCI, PICCOLO, and CPDC dataset 229

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

_____

Signature of candidate

Acknowledgements

First and foremost, I would like to thank Allah Almighty for granting me the strength, knowledge, and perseverance to complete this dissertation. Without His endless blessings and guidance, this achievement would not have been possible.
I am profoundly grateful to my supervisors, Dr. Wei Qi Yan and Dr. Muhammad Usman, for their exceptional guidance and support throughout my doctoral studies. Their expertise, insightful feedback, and dedication have been invaluable in shaping this dissertation.
I would like to extend my deepest gratitude to my parents. To my mother, Rubina Younas, and my father, Muhammad Younas, your unwavering love, encouragement, and belief in me have been the foundation upon which I have built my academic pursuits. Your sacrifices, guidance, and endless support have been invaluable throughout my education. This achievement is as much yours as it is mine.
A special thanks to my husband, Hamza Basit. Your steadfast support, patience, and understanding have been my pillar of strength throughout this entire process. Your belief in me has been a constant source of motivation, and your love and encouragement have helped me persevere through the toughest moments.
A heartfelt thanks go to my wonderful siblings for always being a source of strength for me and for instilling in me the values of hard work, perseverance, and dedication.
Finally, I would like to thank HARVARD Dataverse for making the Colonoscopy Polyp Detection and Classification dataset accessible for this research. I would also like to thank Basque Biobank who generously provided us access to their PICCOLO RGB/NBI database. Their contribution has been instrumental in the analysis and findings presented in this thesis. The support of these organizations has been valuable to the completion of this work.

List of Abbreviations

ARI	**A**djusted **R**ank **I**ndex
cDCGAN	**C**onditional **D**eep **C**onvolutional **G**enerative **A**dversarial **N**etwork
CPDC	**C**olonoscopy of **P**olyp **D**etection and **C**lassification **D**ataset
CRC	**C**olorectal and **C**ancer **C**lassification
DCGAN	**D**eep **C**onvolutional **G**enerative **A**dversarial **N**etwork
DFEB	**D**eep **F**eature **E**xtraction **B**lock
DNN	**D**eep **N**eural **N**etwork
ELB	**E**nsemble **L**earning **B**lock
FID	**F**réchet **I**nception **D**istance
FMI	**F**owlkes **M**ellows **I**ndex
GELU	**G**aussian **E**rror **L**inear **U**nit

IS	**I**nception **S**core
LLE	**L**ocally **L**inear **E**mbedding
MCC	**M**atthew **C**orrelation **C**oefficient
MEB	**M**odel **E**xplaiable **B**lock
MHA	**M**ulti-head **A**ttention
MI	**M**utual **I**nformation
MLP	**M**ulti **L**ayer **P**erceptron
MRB	**M**anifold **R**egularization **B**lock
MSA	**M**ulti-head **S**elf **A**ttention
NMI	**N**ormalized **M**utual **I**nformation
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
SGDM	**S**tochastic **G**radient**D**escent
SIGB	**S**ynthetic **I**mage **G**eneration **B**lock
SSIM	**S**tructural **S**imilarity **I**ndex
UCI	**U**niversity of **C**alifornia **I**rvine
XAI	**E**xplainable **A**rtificial **I**ntelligence

Chapter 1

Introduction

This chapter brings forward the broader context, research aims and contributions, rationale and research methodology of this thesis. It presents a Computer-aided Diagnostic framework for CRC polyp classification. This thesis details that the proposed method results in better classification performance when all challenges posed by biomedical datasets are handled simultaneously. Section 1.1 demonstrates the need for the work done in this thesis. Section 1.2 illustrates the research challenges associated with the data and demonstrates how to handle them. Sections 1.3 and 1.4 list this thesis's research questions and aims and objectives. Section 1.5 elaborates on the research methodology adopted in this study. Lastly, the structure of the thesis is given in Section 1.6, and a list of publications from the thesis is given in Section 1.7.

1.1 Background

Over the last few decades, cancer has emerged as one of the primary causes of death globally. One of the most significant types is Colorectal Cancer (CRC) due to its severe effects and high rates of occurrence and mortality, mostly in developed countries (Pacal et al., 2020). The second highest cause of cancer deaths worldwide is (CRC), which is a serious concern for public health. It also ranks as the third-most prevalent type of cancer diagnosed in general (Jha et al., 2021). In both genders, men and women, colon/rectum cancers are third on the list of new cases in 2021 (GLOBOCAN estimates), with 79,520 men diagnosed with CRC and 69,983 females diagnosed with this disease; these figures represent approximately 8% of all new malignancies among males and females annually. Meanwhile, numbers from only last year show that there were 28,520 U.S. citizens fatally suffering from CRC, whereby men made up 24,460 while women constituted only 4,060 fatalities (Siegel et al., 2021). These stark figures underscore the urgent need for effective diagnostic and treatment approaches targeting CRCs.

Accurate detection and classification of polyps in their early, treatable stages are crucial for reducing the risk of colorectal cancer and preventing associated deaths. Identifying and removing precancerous lesions can significantly prevent the progression to cancer and markedly improve survival rates Levin et al. (2008). Regular screening, particularly in areas with a high prevalence of colorectal cancer, allows for early detection and intervention, which can be instrumental in preventing symptoms from developing and ensuring that tumors do not reach malignant stages. Effective screening programs can thus facilitate timely treatment, reducing the likelihood of severe disease progression. Colonoscopy is widely recognised as the most effective method for detecting colonic polyps, a fact supported by extensive research (Gschwantler et al., 2002; Tajbakhsh et

al., 2016; Shin et al., 2018; Kopelman et al., 2019; Misawa et al., 2018). The procedure allows for direct visualisation of the colon and rectum, enabling the identification and removal of polyps before they can progress to cancer. Routine colonoscopy screening is essential in high-risk populations and is a cornerstone of preventive care in colorectal cancer. The adoption of advanced imaging techniques and improved diagnostic tools further enhances the ability to detect and classify polyps accurately, thereby contributing to better patient outcomes and reduced cancer incidence.

In the primary stages of polyp growth, most polyps are non-cancerous. Removing these polyps is an efficient way to prevent colorectal cancer from advancing. However, a significant number of precancerous polyps can remain undetected during colonoscopy procedures. In such situations, these polyps can develop into cancer at a later stage, thus contributing to increased mortality rates associated with colorectal cancer (Shin et al., 2018). As benign precursor lesions may undergo malignant transformation and lead to cancer-related death, it is crucial not only to identify them but also to classify them reliably. Accurately classifying polyps is important in identifying those who have undergone malignant transformation or are at high risk of doing so. It is not enough to discover some polyps; precise classification should provide insight into whether they can become malignant. This will ensure appropriate interventions that considerably reduce the level of deaths related to colorectal cancer as well as improve patient outcomes through the different approaches used in the management of CRC cases.

1.1.1 Cancer Detection and Classification

In routine medical practice, the removal of polyps varies depending on their size and type. Small polyps are typically removed using techniques such as snares, fulguration,

or biopsy tools. For larger or pedunculated polyps, the snare polypectomy procedure is often employed. This procedure involves using a colonoscope equipped with a wire loop that applies an electrical current to excise the polyp from the mucous membrane of the colon. The primary objective of this procedure is to prepare tumor tissue samples for biopsy, which is crucial for accurate diagnosis and classification of various types of polyps (Geboes et al., 2013). Preparing and examining biopsy slides is notably time-intensive. This involves carefully mounting the tissue samples onto glass slides and conducting a thorough visual examination to identify and classify the polyp types. The accuracy of this classification heavily relies on the expertise and experience of pathologists, making it a critical factor in the diagnostic process. Recent advancements in Medical Image Analysis and Deep Learning (DL) have introduced innovative methods for computer-aided pathology diagnosis, especially in the context of biopsy analysis (Sohail et al., 2021). These technologies leverage sophisticated algorithms and models to assist pathologists in analysing biopsy samples more efficiently. Integrating DL techniques can enhance the accuracy and pace of polyp classification, thereby improving diagnostic outcomes and streamlining the overall pathology workflow.

Integrating machine learning with medical image processing has significantly advanced the identification and classification of cancerous conditions (Litjens et al., 2017). Sophisticated algorithms and methodologies have been developed to enhance Computer-Aided Diagnosis (CAD), enabling more accurate and efficient medical diagnostics (Pacal et al., 2020). Over recent years, AI and DL have become pivotal in analysing and interpreting medical images (Kim et al., 2018; Shen et al., 2017; Thany, 2010). These technologies have revolutionised the field by providing advanced tools for image analysis, which greatly assist in detecting and diagnosing various cancers.

One notable advancement is the significant increase in adenoma detection rates achieved through AI systems. By utilising these artificially intelligent systems, medical image interpretation has evolved, allowing computer-aided diagnosis to function effectively as a secondary layer of analysis for physicians. This support system enhances the diagnostic accuracy and efficiency of medical practitioners. In particular, AI models have proven highly effective in detecting colorectal polyps by interpreting endoscopy images. These models leverage advanced algorithms to analyse and classify polyp images, thereby assisting in early detection and diagnosis (Liew et al., 2021). The application of AI in this context improves the precision of polyp detection and supports the overall goal of enhancing patient outcomes by identifying potential issues at earlier stages.

The effectiveness of diagnosing colorectal lto improve patient outcomes significantlyction Rate (ADR), which is defined as the proportion of colonoscopies that identify at least one adenomatous lesion (Kaminski et al., 2017). Studies have demonstrated that a higher ADR is associated with lower interval rates of colorectal cancer (CRC) and reduced rates of colon cancer mortality (Kaminski et al., 2010; Corley et al., 2014). Flat or sessile polyps, in particular, are often overlooked compared to their larger pedunculated counterparts, which are more easily detected (Hee et al., 2017). Various approaches can be employed to enhance the ADR. An effective method is to use advanced imaging modalities, such as narrow-band imaging (NBI) or white light (WL) endoscopy, to achieve an effective method. NBI, for example, is a sophisticated imaging technique that improves the visualisation of the mucosal surface and accentuates the capillary patterns of polyps. This enhanced visibility facilitates better detection and classification of polyps (Ishaq et al., 2017). By incorporating such advances into CAD systems, there is a potential to significantly improve patient outcomes. NBI's ability

to provide detailed images of the mucosal surface of the colon makes it particularly useful for detecting subtle lesions that might otherwise be missed with conventional imaging techniques. When integrated with advanced CAD systems, this technological enhancement can lead to more accurate diagnoses and potentially better management of risks of colorectal cancer. Furthermore, advancements in CAD systems incorporating such imaging techniques will likely continue improving the overall efficacy of colorectal cancer screening and diagnosis.

Previous CAD algorithms heavily relied on manual feature extraction, significantly limiting their effectiveness in visual object recognition. These algorithms depended on predefined visual features, which constrained their ability to adapt and improve (Bengio et al., 2007). The introduction of deep learning has marked a significant advancement in the field, overcoming the limitations of manual feature extraction by utilising convolutional operations within hidden layers of neural networks. This shift allows deep learning models to automatically learn and extract features from raw data, bypassing the need for explicit manual feature extraction and transformation (Yongzhen et al., 2013). Deep learning has surpassed traditional machine learning techniques and outperformed human visual capabilities in various applications. For instance, DL has been successfully employed in the medical field to automate disease detection and enable early diagnosis. These advanced algorithms leverage complex neural network architectures to process and analyse medical images with high precision, offering substantial improvements over conventional methods (Sahiner et al., 2019). This progression demonstrates the transformative impact of deep learning on CAD systems, providing more accurate and efficient tools to diagnose and manage diseases.

Integrating a CAD system into the manual biopsy process is crucial to enhancing

diagnostic efficiency and accuracy. This integration can streamline the workflow, assist pathologists by providing additional decision support, and potentially reduce diagnostic errors. A virtual biopsy CAD system can facilitate more accurate assessments by analysing digital images and providing predictive insights based on advanced algorithms. However, to ensure the effectiveness and reliability of these systems, a thorough evaluation of their clinical viability is essential. This includes evaluating their performance in real-world settings, assessing their impact on diagnostic accuracy, and understanding their integration into clinical workflows. Such evaluations help determine whether these systems can effectively complement traditional methods and improve patient outcomes. By rigorously testing and validating CAD systems, healthcare professionals can ensure that these technologies are effective, safe, and beneficial for clinical use. (Suzuki, 2012).

1.1.2 Deep Learning in Cancer Classification

In recent years, the fields of artificial intelligence (AI) and DL have made notable strides in the realm of medical image analysis, leading to significant advancements over traditional machine learning (ML) approaches (Kim et al., 2018) (Chan et al., 2020). Unlike conventional ML methods, which typically rely on an explicit definition of features and substantial manual intervention, DL algorithms stand out by their ability to autonomously process and integrate high-dimensional data features. This automation negates the necessity for manual feature extraction, thereby enhancing efficiency and performance. The superior capabilities of DL have enabled these techniques to outshine traditional ML methods, such as support vector machines (SVM) and random forests. These older methods are increasingly viewed as outdated in comparison because of their limitations. Traditional ML models often require extensive manual design and

development, rendering them less robust, time-consuming, and adaptable to new data. The real-time applicability and overall effectiveness of these older techniques fall short when juxtaposed with the advancements achieved through DL methods. DL algorithms have demonstrated considerable improvements in analysing medical images, such as colon images, and in detecting abnormalities with greater accuracy and efficiency. These advancements highlight a significant leap forward from traditional ML models, which struggle to achieve similar levels of performance and precision (Sohail et al., 2021). The enhanced capabilities of DL not only streamline the analysis process but also contribute to more reliable and timely medical diagnoses.

Deep learning computer-aided diagnosis (DL-CAD) has emerged as a notably effective and comprehensive approach for cancer diagnosis, showcasing its transformative impact on medical imaging and analysis (Komeda et al., 2017). Among the various deep learning methods, convolutional neural networks (CNNs) stand out as a foundational component, and they are celebrated for their exceptional performance in classification tasks. CNNs have become instrumental in medical image analysis due to their remarkable ability to classify and interpret complex image data accurately. These sophisticated algorithms are extensively utilised across various colonoscopy applications, including but not limited to the detection, classification, segmentation, and localisation of abnormalities within the colon. However, despite the widespread application of deep learning techniques for the detection of polyps, there remains a notable lack of emphasis on the precise classification of different types of polyps (Sanchez-Peralta et al., 2020). This gap can be largely attributed to several significant challenges, including the limited availability of large-scale, annotated medical datasets that are crucial for training advanced models capable of achieving high levels of classification accuracy (Nogueira-Rodríguez et al.,

2021). These annotated datasets are essential for developing and refining sophisticated models that can accurately differentiate between various types of polyps, which is vital for effective treatment planning and optimal patient management. The current state of progress reveals that while there have been substantial advancements in the detection of polyps, the field of accurate polyp classification has not kept pace. This disparity underscores a critical need for ongoing research and development efforts to improve classification capabilities. Addressing the challenges associated with the scarcity of comprehensive medical datasets is essential for bridging this gap and advancing the field of DL-CAD. Enhanced classification abilities will improve diagnostic accuracy and contribute significantly to more personalised and effective treatment strategies (Pacal et al., 2020).

However, practical applications of deep learning-based systems for endoscopic detection and classification continue to tackle many significant challenges. Despite the considerable technological advancements in this field, these systems frequently display a degree of unreliability when deployed in clinical settings. This raises important concerns about the effectiveness and consistency of such systems in real-world medical environments (Kudo et al., 2019). The integration of deep learning into colonoscopy procedures, while holding substantial promise for revolutionising endoscopic practices, still encounters hurdles that impact its practical implementation. Deep learning has the potential to greatly enhance the standardisation of endoscopic procedures, which could lead to a reduction in medical errors and support specialists in making more accurate and reliable diagnoses. Improved diagnostic accuracy is crucial in the medical field, where precision and dependability are of utmost importance. By providing a more consistent and automated approach to analysing endoscopic images, deep learning systems could

transform the way diagnoses are made, potentially leading to better patient outcomes and more effective treatment strategies. Nevertheless, the current state of deep learning technology in this area underscores the need for ongoing research and development. Addressing the limitations and shortcomings of these systems is essential to ensure that they can be effectively trusted and implemented in real-world clinical settings. Efforts to refine these technologies and overcome their current challenges are critical for realising the full potential of deep learning in endoscopic diagnostics. As these systems evolve, achieving greater reliability and effectiveness will be key to their successful integration into routine medical practice.

1.2 Research Problems

Two major factors that will contribute to the further advancement of deep learning in endoscopy include the accessibility of high-quality endoscopy images and the improvement of endoscopist decision-making abilities (Min et al., 2019).

1.2.1 Data Accessibility

Although deep neural networks (DNNs) have demonstrated high accuracy in classifying colorectal polyps from endoscopy images, their application in virtual biopsy is actively explored to automate the pathologist's workflow and enhance their capabilities by developing sophisticated classifiers. A critical factor driving the development of effective deep learning models is the availability of large, high-quality datasets (Y. Zhang et al., 2023). Despite this, there is a notable scarcity of publicly accessible datasets specifically for colorectal polyps, which is compounded by the issue of class imbalance. Obtaining and annotating medical data is expensive and labour-intensive, posing a significant

challenge in creating a comprehensive, well-annotated training dataset (Jha et al., 2021; Nogueira-Rodríguez et al., 2021). Transfer learning and data augmentation are employed to address these challenges. Transfer learning leverages pre-trained models, which have been trained on extensive datasets, by adapting their learned weights for new, related tasks. This method is particularly useful when dealing with limited data, as it allows the model to benefit from the knowledge acquired from large-scale datasets. However, transfer learning was not utilised in our approach because the available pre-trained models were not based on biomedical images. Consequently, the reliability of the model, when based on weights captured through non-biomedical data training, may be compromised. This limitation underscores the need for more specific and relevant biomedical datasets to harness the potential of transfer learning fully in the context of medical image analysis.

Our research addresses the issues associated with limited data and class imbalance by employing advanced data augmentation techniques, which are crucial for enhancing various classification tasks (Vanyan & Khachatrian, 2021). Specifically, we focus on generating unseen and reliable synthetic images to increase the volume and diversity of the training data. This augmentation improves the generalizability of the classification model by introducing new variations of images from which the model can learn. However, these synthetic images must exhibit high concordance with real data to effectively contribute to building an intelligent classifier. To tackle this challenge, we have integrated Conditional Generative Adversarial Networks (cGANs) into our framework to produce synthetic images. Generative Adversarial Networks (GANs) are known for introducing new, unseen images into the dataset, enhancing data diversity. On the other hand, cGANs extend this capability by focusing on generating images of specific class labels with the assistance of additional information (Krizhevsky et al.,

2012; Sener & Savarese, 2017). This feature is useful for addressing the class imbalance problem prevalent in colorectal cancer (CRC) datasets. Despite the benefits, it is essential to rigorously evaluate both the quality and diversity of the synthetic images generated by the cGAN network. To ensure that the synthetic images benefit training, we must assess their concordance with real images before integrating them into the training dataset. Without this evaluation, the classifier's performance might deteriorate due to introducing less relevant or noisy synthetic images. Our proposed framework addresses this by incorporating an additional classifier that separates the synthetic images based on their predicted class probability. This classifier ensures that only highly concordant images are used for training, which helps maintain the final classifier's performance. We assess the quality and diversity of the synthetic images using four effective measures: FID - Fréchet (Heusel et al., 2017), IS - Inception Score (Park et al., 2022), PSNR - Peak Signal-to-Noise Ratio Wang et al. (2004) and SSIM - Structural Similarity Index Wang et al. (2004). These metrics provide valuable insights into the fidelity and diversity of the generated images, ensuring that they contribute positively to the training process.

Several strategies are employed to enhance the performance of classifiers by utilising diverse augmentation techniques. Among these, GANs have gained significant popularity for data augmentation purposes. GANs are deep learning models consisting of two neural networks, the generator and the discriminator, which compete against each other to produce realistic images (Elsheikh, 2008). The generator creates synthetic images, while the discriminator evaluates their authenticity, generating increasingly realistic images through this adversarial process. Conditional Generative Adversarial Networks (cGANs), an extension of GANs, introduce additional information to control the generation process more precisely, allowing the creation of images for specific classes

(Krizhevsky et al., 2012; Sener & Savarese, 2017).

cGANs' ability to focus on particular class labels enhances the diversity of training data and helps address class imbalance issues. By generating synthetic images representing various classes, cGANs significantly improve the model's ability to learn from a more varied dataset.

In addition to GANs and cGANs, Vision Transformers (ViT) have shown remarkable effectiveness in capturing long-range dependencies within image data, making them adept at handling complex contextual information. ViT models are designed to process large images and extract detailed features, including fine-grain details, crucial for high-quality image analysis. The ability of ViT to handle complex image structures complements the strengths of GANs and cGANs. Combining these techniques, GANs for data augmentation and ViT for feature extraction offer a robust approach to addressing the challenges of data insufficiency and class imbalance. This integrated strategy enhances the model performance by increasing the diversity of training data and improving the extraction of meaningful features from images.

1.2.2 Feature Extraction and Model Transparency

Feature extraction plays a crucial role in the analysis of medical images, as it directly impacts the accuracy and effectiveness of diagnostic models. Currently, the methodologies used to extract the characteristics of colorectal polyps include mainly end-to-end learning, hand-made characteristics, and hybrid methods (Bernal et al., 2017). Each of these approaches has distinct advantages and limitations in capturing the relevant characteristics of polyps. End-to-end learning involves training models, particularly deep learning networks, to learn features directly from raw image data without manual

intervention. This approach leverages neural networks to automatically discover and extract relevant features from images, eliminating the need for explicit feature engineering. On the other hand, hand-crafted feature extraction relies on manually designed algorithms to identify and quantify specific attributes such as texture, shape, colour, and intensity. These features are engineered based on domain knowledge and are used to train traditional machine learning models. Although hand-crafted methods offer a high level of control over feature selection, they can be limited by the designer's expertise. They may not capture all relevant aspects of the data.

Hybrid methods combine end-to-end learning and hand-crafted approaches, integrating the strengths of each to improve feature extraction. For instance, these methods might use deep learning to automatically extract high-level features while incorporating manually designed features to capture specific details that are particularly important for accurate classification. Despite these advancements, it remains crucial to extract complete and accurate features, including fine-grain details, to minimise the misclassification rate of polyps. The quality and granularity of the extracted features significantly influence the performance of the classification models. Inaccurate or incomplete feature extraction can lead to misclassification, affecting the reliability of diagnostic results. Therefore, ongoing research and development efforts focus on refining feature extraction techniques to enhance the precision and effectiveness of colorectal polyp detection and classification systems.

The sourcing of data and the relevant features have been the two determining factors influencing the prediction accuracy of CAD systems. Traditionally, it has been observed that those trying to resolve these problems, especially in the biomedical domain, have focused on one of the two aspects at a time, either increasing the data availability or

enhancing the feature extraction approaches. Nevertheless, completing the last two steps will be dealt with simultaneously, which is complicated in several aspects. Firstly, data augmentation methods are widely used to address the issue of limited data. Common practices include rotation, flipping, adjusting brightness, and scaling of images. While these techniques are effective for increasing the size of the dataset, they do not introduce new or unseen samples, limiting the training data's diversity. As a result, the dataset may still lack sufficient variability to effectively train models, leading to potential limitations in the model's ability to generalise to new, real-world data. Secondly, the reliance on traditional data augmentation methods also impacts the feature extraction process. Since these methods do not introduce novel images but rather modify existing ones, they often fail to capture the full range of variability in real-world scenarios. This limitation results in a dataset that may not fully represent the diverse features necessary for accurate classification. Consequently, the feature extraction process is constrained, decreasing classifier sensitivity and performance. The model may struggle to differentiate between subtle feature variations, which can negatively impact its ability to classify and predict outcomes accurately. Therefore, addressing data availability and feature extraction challenges is crucial for developing robust CAD systems. Innovative approaches that combine advanced data generation methods with sophisticated feature extraction methods are needed to overcome these limitations. This includes incorporating strategies such as synthetic data generation, leveraging new augmentation methods, and improving the extraction of detailed and discriminative features to enhance CAD systems' overall accuracy and reliability.

This research aims to resolve challenges in the feature extraction process by

incorporating the Vision Transformer (ViT) model into the framework. The Vision Transformer, introduced by (Dosovitskiy et al., 2020), represents a significant advancement in image classification and has shown remarkable efficiency in feature extraction tasks. The ViT model leverages a novel approach inspired by how humans perceive and analyse images. Just as humans focus on specific regions of an image to identify areas of interest, ViT applies a similar concept to image classification, allowing it to efficiently capture and process relevant features (Han et al., 2022).

The Transformer architecture, originally designed for Natural Language Processing (NLP) tasks (Vaswani et al., 2017), has achieved substantial success in understanding and generating human language. Its ability to handle complex dependencies and contextual information in text has sparked interest among computer vision researchers in exploring whether this architecture could also excel in visual tasks. This curiosity has led to a growing body of research investigating the potential of Transformer models as strong contenders to traditional Convolutional Neural Network (CNN) architectures, such as ResNet, GoogLeNet, and others. CNNs have long been the dominant approach for image analysis due to their ability to capture spatial hierarchies and local patterns in visual data effectively. However, the advent of Transformer models, with their ability to model long-range dependencies and capture global context, has prompted a reevaluation of their effectiveness compared to CNNs. The exploration of ViT and other Transformer-based models in computer vision is driven by the potential for these models to provide superior performance in feature extraction and image classification tasks, particularly in complex and high-dimensional data scenarios.

In recent years, research on transformers in computer vision has been primarily focused on integrating Convolutional Neural Networks (CNNs) with self-attention

mechanisms (Bello et al., 2019; Zhao et al., 2020; Ramachandran et al., 2019; Srinivas et al., 2021). While these hybrid approaches have shown promising results, they often fall short in terms of computational scalability when compared to purely attention-based transformers. The Vision Transformer (ViT) represents a significant advancement in this area. It is the first pure attention-based model demonstrating performance on par with traditional CNNs. The ViT operates by taking a sequence of embedded image patches as input, aligning with the principles of the standard transformer architecture (Vaswani et al., 2017). However, one of the notable limitations of the ViT model is its requirement for a substantial amount of training data to achieve optimal performance. For instance, it typically requires datasets as extensive as ImageNet21k (Deng et al., 2009) to leverage its full potential. Despite this, the ViT model excels in learning and extracting image features. The final layer of the ViT model can effectively capture and extract meaningful features, which can then be utilised for further tasks. Integrating conditional GANs (cGANs) into the framework can be highly beneficial in addressing the challenge of dataset limitations. cGANs can generate synthetic images to augment the dataset, fulfilling a large training dataset requirement and enhancing the model's overall performance.

While newer transformer-based architectures, such as Swin Transformer, DeiT, Hybrid Transformers, and PVT (Pyramid Vision Transformers), offer computational efficiency, they primarily focus on capturing local relationships in image patches. This design makes them well-suited for scalability and small datasets but less effective in medical imaging tasks requiring distinguishing subtle differences between anatomical structures. In contrast, Vision Transformers (ViTs) leverage a self-attention mechanism that captures long-range dependencies and global relationships, which is crucial

for precise medical image classification—especially for tasks like differentiating polyp boundaries from surrounding tissues in colorectal cancer detection. Many transformer variants prioritise computational efficiency over feature richness, leading to challenges in handling complex medical images. This trade-off can significantly impact classification accuracy in high-risk applications, where misclassification can have serious clinical consequences. Additionally, some variants introduce architectural complexities, such as shifting windows in Swin Transformers, which may increase model interpretability challenges and computational overhead without improving feature extraction quality. One of the key distinctions between ViTs and their variants is the treatment of global receptive fields. Swin Transformers, for instance, utilise local attention windows to reduce computational cost, making them more scalable for high-resolution images. However, this comes at the cost of a reduced global receptive field, crucial for identifying structural relationships in medical imaging. While introducing shifting windows helps somewhat mitigate this limitation, adding additional architectural complexity may not necessarily translate to better feature extraction. ViTs, on the other hand, maintain a global receptive field from the beginning, enabling them to capture contextual relationships across the entire image. This is particularly important in medical imaging tasks where subtle differences—such as a polyp's faint boundary against the colon wall—can decide between a correct or incorrect diagnosis. While DeiT (Data-efficient Image Transformers) is optimised for small datasets using knowledge distillation techniques, its performance on large-scale medical datasets may not match standard ViTs. ViTs' global attention mechanism excels at learning complex patterns in large datasets, making them more suitable for medical applications involving vast amounts of high-resolution data. Architectures like Swin Transformer and PVT trade off feature richness for computational efficiency. While they efficiently model local spatial relationships, their limited ability to capture

fine-grained, long-range dependencies can be problematic in medical image classification, where minor differences in texture, shape, or contrast determine clinical decisions. In high-stakes environments such as cancer diagnosis, compromising on feature extraction is not acceptable, as errors can lead to misdiagnoses and potentially life-threatening consequences. While newer transformer architectures provide computational benefits, their feature extraction and long-range dependency modeling trade-offs make them less ideal for medical image analysis, particularly in colorectal polyp detection. Given the sensitivity of medical data, ViTs remain a strong choice due to their ability to capture global relationships without sacrificing interpretability or feature richness. Future advancements should aim to balance efficiency with diagnostic accuracy, ensuring that computational optimisations do not come at the cost of patient safety.

Another significant challenge in deep learning models is their inherently opaque decision-making process. Automated diagnosis systems must address this problem to build trust and reliability with medical practitioners. Developing an explainable CRC classification system for colonoscopy imaging is essential to bridge this gap effectively. Such a system should provide insight into the features that contribute to the final predictions of the model, thereby assisting pathologists in making well-informed decisions. Our research framework is designed to deliver a deep learning-based classification model that achieves high performance and maintains transparency and interpretability. To realise this objective, we leverage the feature extraction capabilities of the Vision Transformer (ViT) model. The features extracted by the ViT model are then visualised using Grad-CAM (Gradient-weighted Class Activation Mapping). Grad-CAM is a visualisation technique that generates heatmaps to highlight areas of an image that significantly influence the decision-making process of the model. This method allows for a detailed representation

of the focus areas of the model, providing valuable information on how different characteristics contribute to the final classification. By incorporating such explainable AI techniques, our framework aims to enhance the interpretability of deep learning models in CRC classification, ultimately fostering greater trust and confidence among pathologists and healthcare professionals.

1.2.3 Nonlinear Dimensionality Reduction and Model Regularisation

An increased number of extracted features often results in high-dimensional data, which poses a significant challenge to capture meaningful patterns among the dispersed data points. This complexity can lead to notorious overfitting problems, where the model performs well on training data but poorly on unseen data. Addressing this issue is crucial for building a generalised and robust model, as overfitting can drastically affect the model's performance and reliability. Several strategies are employed to mitigate the problem of overfitting, such as early stopping or regularisation of the parameters (Ying, 2019). However, these methods can sometimes limit the maximum representational capacity of the trainable filters, potentially reducing the model's ability to capture intricate patterns in the data. Images, especially those used in medical diagnostics, often contain complex, nonlinear structures that linear methods alone cannot effectively represent. This limitation necessitates the use of advanced techniques to enhance feature representation. In this context, nonlinear dimensionality reduction techniques are more appropriate for our data. These methods are designed to uncover and preserve the underlying structure of high-dimensional data in a lower-dimensional space, which is particularly useful for dealing with complex and nonlinear patterns in image data. Our framework incorporates regularising the manifold algorithm's cost function to improve feature extraction and

representation. This approach aims to address the challenge of high-dimensional data by leveraging the manifold structure of the data, thereby enhancing the accuracy and generalisability of the model. By integrating such advanced techniques, we strive to build a more effective and robust model for analysing and classifying medical images.

The manifold algorithm aims to represent the complexity in the geometry of images, which are represented as high-dimensional data in a way that will be encoded in low-dimensional embeddings that are formed and subsequently used. The goal of manifold learning for dimensionality reduction is to identify a mapping from a high-dimensional manifold to a lower-dimensional space, allowing for the representation of data points in the high-dimensional space while effectively reducing the dimensionality of the data (WEI, 2023). This activity is essential since most of the image information is complex and, most of the time, has underlying structures that are not linear and cannot be modelled traditionally. Several dimensionality reduction techniques are usually employed to solve these problems. The most common are multidimensional scaling (MDS) and principal component analysis (PCA). MDS and PCA techniques are both linear transformations that allow for decreasing the dimension of the supplied data. However, such linearity can hinder the learning and modelling of nonlinear manifolds. Linearity does not always work well when nonlinear relationships are in the form of high-dimensional data, resulting in inadequate feature abstraction.

In contrast, these non-linear dimensionality reduction approaches target high-dimensional space structures more efficiently by finding and retaining appropriate geometry. For example, Isometric Mapping, or Isomap, has the property that distances are preserved but maps the data points undistorted in a less dimensional space. This method does an excellent job of preserving the geometric arrangement of the data's internal

structure to create complicated structures in lower dimensions.

Unlike Locally Linear Embedding (LLE), which preserves local neighbourhoods' structures well, it guarantees the degradation of global neighbourhoods' hierarchy. LLE is more efficient than its other counterparts in exploring the data structure because it puts more attention on the local structure of the manifold, which is important in applications that need to understand the detailed structure and subtle changes of the data. These types of nonlinear dimensionality reduction techniques have considerable advantages over linear methods. They help to understand the much deeper and more sophisticated structure of the higher-dimensional data. To take advantage of these methods, the proposed approach intends to leverage their benefits in improving the effectiveness of the feature extraction process and, consequently, the model's overall performance in the analysis and classification of image data.

1.2.4 Model Evaluation

Medical risks brought about by automated diagnosis systems are unlikely to be acceptable, especially in CRC. Thus, imaging and classification of polyps are necessary for effective diagnosis. Overlooking polyps can create several problems, including delays in critical polypectomy operations and inappropriate decisions related to further resection. Such delays or errors may be risky for complete risk neoplasia progression and ultimately have fatal consequences for the patient. To mitigate these risks and ensure the reliability of automated diagnostic systems, it is essential to incorporate a comprehensive set of advanced evaluation metrics beyond basic accuracy measures. Precision, recall, and F1 score are critical metrics that provide a more nuanced understanding of the model's performance. Precision measures the proportion of true positive classifications among

all positive identifications made by the model, while recall assesses the proportion of actual positives that were correctly identified. The F1 score, precision, and recall offer a balanced evaluation of the model's ability to correctly identify relevant cases while minimising false positives and negatives.

In simpler terms, addressing type II classification errors, also known as false negatives, is essential to ensure the effectiveness of diagnostic tests. A high-sensitivity test is crucial because it guarantees that all identified negatives are truly negative, which means the test correctly identifies every instance where the condition is absent. This high sensitivity is particularly valuable for ruling out diseases and is crucial for screening tests, especially for conditions with low prevalence where missing a case could have severe consequences.

Colonoscopy is arguably the most sensitive diagnostic procedure to limit colon cancer and strengthens colon cancer screening efforts. Potential cases of colon cancer are not missed due to the high level of sensitivity, making it easier to initiate care and treatment. This is due to the importance of colon cancer, where high-sensitivity tests should seek to eliminate false negatives that can present and delay the diagnosis of patients. Performance evaluation, as a result, in this case, focuses on recall measures. Recall or sensitivity, in this case, is important because it measures how well the test identifies people with positive cases. In this case, incorrect classification of malignant polyps would lead to a high false negative rate, which would be detrimental to patients. In this case, a high false negative rate means that malignant polyps would be missed, resulting in the development of cancer, which in turn is unfavourable for the patient. Hence, it is important to incorporate high sensitivity into the diagnostic architecture to avoid misdiagnosis and achieve adequate cancer screening.

1.3 Research Questions

RQ1: Can an efficient deep learning colorectal polyp classification model be developed using advanced data augmentation techniques for insufficient training data?

Sub-questions:

1a. Which set of hyperparameter settings contributes to optimal performance in building a CRC polyp classifier using transfer learning?

1b. Which is the best-fit deep learning model to successfully assist endoscopists in performing virtual biopsies?

1c. Which data augmentation technique is most effective for enhancing the quality and diversity of the CRC polyp dataset?

RQ2: How can deep learning models be made more interpretable and transparent for CRC classification, ensuring that pathologists can trust and understand the decision-making process?

Sub-question:

2a. How can we extract and visualise key features in CRC polyp images that contribute to its predictions?

RQ3: Can nonlinear dimensionality techniques regularise the CRC classification model and achieve a successful classification task?

Sub-question:

3a. To what extent can nonlinear dimensionality reduction techniques such as Isomap, t-SNE, or LLE simplify the feature space complexity and improve

the regularisation of the CRC classification model?

RQ4: Does the CNN classification model performance further enhance by developing custom-built networks for accurate colorectal polyp classification?

Sub-question:

4a. How do custom-built CNNs compare to pre-trained models using transfer learning for colorectal polyp classification?

1.4 Research Aims and Contributions

Our research aims to accomplish the following objectives:

- To provide a DL-powered solution for CRC virtual biopsy to help pathologists distinguish between those that should go under histopathology and those that should be removed directly.
- To assist in diagnosing colonic polyps with minimal training for endoscopists through deep learning-assisted colonoscopy.
- To standardise endoscopy procedures and reduce human error while assisting experts in improving accuracy and sensitivity of diagnoses.
- To handle the problem of data insufficiency and class imbalance by generating high-quality and diverse synthetic CRC image data.
- To provide an explainable deep learning model essential for pathologists in interpreting the decision-making process.
- To discover significant patterns within complex high-dimensional images and to

regularise the deep learning algorithm for better generalisation and to prevent overfitting.

We have contributed the following main points to this thesis:

- Introduced an innovative framework for transfer learning-based virtual biopsy, focusing on classifying colorectal polyps captured under NBI lighting. The framework addresses the challenge of limited dataset images by leveraging transfer learning and image augmentation.

- Evaluated and compared the efficacy of various structures in deep learning, namely GoogLeNet, ResNet50, Inception-v3, Xception, DenseNet-201 and SqueezeNet, to identify the most suitable deep neural network for the classification of colorectal polyps with traditional data augmentation techniques and determining optimal hyperparameter configurations to optimise deep neural networks.

- Proposed a weighted average ensemble model to deal with the complex nature of polyps by improving the classification system's generalisation and classification results through a weighted average ensemble learning technique by assigning weights to base classifiers, contrasting with ensemble averaging and cutting-edge CNNs.

- Proposed a heterogeneous ensemble approach for classifying colorectal polyps in colonoscopy images. The model enhances generalisation and robustness by leveraging the strengths of independent CNNs within the ensemble learning framework.

- Classified polyps into three classes: serrated polyps in addition to hyperplastic and adenomatous polyps, which lead to CRC through an alternate serrated pathway,

which is challenging to identify.

- Presented a novel framework for a concordant conditional generative adversarial classifier that generates realistic and diverse CRC images. Synthetic images generated were used for efficient training, asserting the strength of a unified cDCGAN and classifier network to generate unseen data for training a classifier while enhancing the generalisability of the classifier.

- Investigated the unified ability of the proposed cDCGAN and ViT for effective classifier training on insufficient biomedical image datasets and feature extraction to achieve multiclass CRC polyp classification.

- Presented an explainable deep learning method, XAI, using Grad-CAM to build the trust of pathologists and enhance the transparency of the decision-making process of the developed method.

- Presented the fusion of synthetic data generation and feature extraction with a nonlinear manifold learning technique to alleviate the curse of dimensionality and performed model regularisation.

1.5 Research Methodology

This thesis employs the Design Science Research Method (DSRM) process model, a problem-solving paradigm dedicated to advancing human knowledge by creating innovative and effective artefacts. The DSRM framework is particularly suited for addressing complex challenges by designing and evaluating solutions that provide practical and theoretical contributions.

In our approach, we conducted a thorough analysis of existing research to identify gaps in the current knowledge base and explore potential areas for improvement. Our motivation for conducting this research is rooted in the pressing need for a reliable virtual biopsy assistance tool specifically tailored for pathologists dealing with colorectal cancer (CRC). Clinicians often encounter difficulties trusting the decision-making processes of deep learning models, especially those developed for medical image classification. This lack of trust is often due to the opaque nature of these models, which hinders their integration into clinical practice.

To address these challenges, our study aims to develop an accurate and efficient deep learning-based computer-aided diagnostic (CAD) tool designed to classify the subtypes of polyps of CRC polyps effectively. Furthermore, we strive to enhance the explainability of the model's predictions, which is crucial for building confidence among clinicians and pathologists. By formulating a deep learning framework, we address the issues of insufficient biomedical datasets and class imbalance. This is achieved by integrating a concordant conditional generative adversarial network (cGAN), which helps generate synthetic images to supplement the existing dataset.

Additionally, the study focuses on developing an explainable model using Vision Transformers (ViT) combined with the Grad-CAM approach. This combination is intended to elucidate the black-box nature of deep learning models, providing insights into how predictions are made. Given the large dataset and the high dimensionality of images, we also incorporated nonlinear manifold learning and model regularisation techniques. These methods are essential for managing the complexities of high-dimensional data and ensuring that the model remains robust and generalisable.

We conducted extensive experiments to thoroughly validate the proposed deep learning framework, utilising it across three distinct CRC polyp datasets. These datasets were acquired from both public repositories and private sources, ensuring a diverse range of data for robust testing. By leveraging data from multiple sources, we aimed to capture various characteristics of CRC polyps, thereby enhancing the generalisability and reliability of our framework. The experimental procedures were meticulously designed to comprehensively assess the effectiveness of our model on each dataset. For each dataset, we conducted multiple rounds of testing, employing a variety of advanced evaluation metrics to gauge the model's performance accurately. These metrics included but were not limited to accuracy, precision, recall, F1-score, and sensitivity. Using a comprehensive set of metrics ensured that the model's strengths and weaknesses were fully understood, allowing us to refine and optimise its performance.

In addition to evaluating the overall performance of the proposed deep learning framework, we also broke down the evaluation process to assess the contribution of each component within the framework. Each framework part was subjected to a separate evaluation using its corresponding evaluation metrics. This granular approach allowed us to identify which components were most effective and which needed further refinement. By analysing each part individually, we could ensure that every aspect of the framework functions optimally and contributes positively to the overall performance.

Furthermore, these experiments provided valuable insights into how different elements of the framework interacted with each other. Understanding these interactions was crucial for fine-tuning the model and ensuring consistent and reliable results across all datasets. The thoroughness of this evaluation process underscores our commitment to developing a deep learning solution that is accurate, highly reliable, and adaptable

to different types of data. This thesis not only seeks to advance the field of medical image classification but also aims to improve the trust and reliability of deep learning models in clinical settings by addressing key challenges related to data insufficiency, class imbalance, and model interpretability. Our comprehensive experimental approach, combined with detailed evaluations of each component, has allowed us to validate the effectiveness of the proposed deep learning framework rigorously. This methodical process has provided a deep understanding of the model's capabilities and limitations, guiding us toward further improvements and ensuring that the final solution is robust and dependable in real-world applications.

1.6 Thesis Structure

To address the research aims specified, this thesis is outlined as follows:

Chapter 2 provides a comprehensive overview of six pivotal themes in our research area, each contributing to the advancement of computer-aided diagnosis (CAD) for colorectal cancer (CRC) and related fields. Section 2.1 delves into previous studies on CRC CAD systems, presenting a detailed examination of various CAD approaches' evolution, methodologies, and effectiveness over time. This section reviews how these systems have been developed and refined, including their successes and limitations in improving diagnostic accuracy and efficiency. Section 2.2 focuses on synthetic image generation and its critical role in enhancing data diversity and quality. This section explores various techniques used to generate synthetic images, evaluating their effectiveness in addressing data insufficiency and variability challenges. The review highlights how synthetic image generation can bridge gaps in data availability and

contribute to more robust training datasets. In Section 2.3, we review advancements and challenges in feature extraction techniques, which are essential for building accurate and reliable classification models. This section provides an in-depth analysis of different methods for extracting features from medical images, discussing their strengths and limitations. It emphasises the importance of extracting detailed and informative features to improve the performance of classification systems. Section 2.4 examines the progress and current practices in model explainability, particularly in the context of medical diagnosis systems. This section discusses how explainability techniques are implemented to make deep learning models more transparent and interpretable. By providing insights into these models' decision-making processes, explainability helps build trust among practitioners and improve diagnostic decision-making. In Section 2.5, we review feature selection methods, highlighting their role in refining model performance and addressing issues such as class imbalance. This section discusses various feature selection techniques and their impact on model accuracy and efficiency, stressing the need for effective feature selection strategies to enhance the overall performance of diagnostic models. Finally, Section 2.6 covers model regularisation and its application in image classification. This section explores different regularisation techniques to improve model generalisation and prevent overfitting. It discusses how these methods contribute to developing more robust and reliable classification models by managing the complexity of the models and ensuring they perform well on unseen data.

Chapter 3 offers a detailed overview of the proposed methodology for colorectal polyp classification, providing a thorough illustration of each methodological step involved. This chapter emphasises how integrating concordant image generation with

advanced feature extraction techniques can facilitate the development of a highly explainable and effective classification model. By leveraging these methodologies, we aim to address various challenges associated with CRC polyp classification and enhance overall model performance. We illustrate how concordant image generation can be combined with feature extraction processes to improve model explainability and accuracy. The chapter discusses the application of nonlinear dimensionality reduction techniques, such as Isomap and LLE, which are instrumental in preserving data's statistical and geometric properties when reduced to lower dimensions. These techniques also play a crucial role in regularising the model, enhancing its classification performance and generalizability. Section 3.1 presents the methodological framework developed in this study, detailing the comprehensive approach employed for CRC polyp classification. This section lays the foundation for understanding how various methodology components are integrated to achieve the desired outcomes. Section 3.1.1 delves into the capabilities of the weighted ensemble method and the importance of hyperparameter optimisation in classification tasks. This section addresses research question **RQ1**, highlighting how these techniques improve model performance and accuracy by combining the strengths of multiple models and fine-tuning hyperparameters. In Section 3.1.2, we discuss enhancing synthetic image generation using conditional adversarial networks to create high-quality, concordant datasets. This section demonstrates how synthetically generated images can be optimised to address data insufficiency issues and improve the classification model's robustness. Section 3.1.3 elaborates on deep feature extraction using attention-based mechanisms, addressing the research question **RQ2**. This section emphasises the importance of extracting fine-grain details from polyp images to enhance the classification model's sensitivity and accuracy. Section 3.1.4 focuses on integrating explainable AI (XAI) into the classification pipeline, which facilitates the understanding of the model's

decision-making process. This section addresses how XAI methods help interpret model predictions and build trust among practitioners. Section 3.1.5 examines the significance of nonlinear dimensionality reduction and model regularisation in biomedical image classification, addressing research question **RQ3**. This section illustrates how these techniques contribute to managing the complexity of high-dimensional data and improving model performance. Finally, Section 3.1.6 presents the final classification results and addresses the research question **RQ4**. This section provides a comprehensive analysis of the classification performance, evaluating the effectiveness of the proposed methodology in achieving accurate and reliable CRC polyp classification.

Chapter 4 provides an in-depth discussion of the experiments conducted using the developed framework, offering a comprehensive analysis of each phase of the experimentation process. Section 2.7 delves into the problem formulation, outlining the specific challenges and objectives addressed by the framework. This section is crucial for understanding the context and rationale behind the experimental design, including formulating research questions and hypotheses that guide the validation of the model. Section 4.1 presents a detailed account of the datasets utilised for training and evaluating the learning model. This section describes the datasets' sources, characteristics, and preprocessing, including any data augmentation or enhancement techniques applied to improve dataset quality and diversity. Section 4.2 details the experimental setup, providing a thorough description of the implementation procedures. This includes the configurations and parameters used during the experiments and the hardware and software environments in which the experiments were conducted. The section offers insights into the practical aspects of executing the experimental protocols. Section 4.3 outlines the network architecture employed in the experimental phase, detailing the specific deep learning models

and configurations used. This section includes information on the network layers, hyperparameters, and any modifications made to the base architectures to suit the experimental requirements. Finally, Section 4.4 elaborates on the advanced evaluation metrics applied to assess the performance of the model. This section explains each metric used, including their relevance and application in evaluating the model's effectiveness. It also discusses how these metrics contribute to a comprehensive understanding of the model's strengths and areas for improvement, ensuring that the model is rigorously tested and validated.

Chapter 5 presents a comprehensive analysis of the results derived from the experiments outlined in the previous chapter, offering detailed insights into each aspect of the evaluation process. Firstly, Section 5.1 focuses on the outcomes of extensive experimentation with hyperparameter tuning, transfer learning, and ensemble learning. This section thoroughly analyses how these techniques influenced the model's performance. The discussion covers the impact of various hyperparameters on model accuracy, the benefits of incorporating transfer learning, and the effectiveness of ensemble learning in enhancing the model's robustness and generalizability. The results are analysed to highlight key findings and implications for future work. Next, Section 5.2 delves into the results of synthetic image generation and deep feature extraction, including an in-depth analysis of evaluation metrics such as Fréchet Inception Distance (FID), Inception Score (IS), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). This section discusses the quality and diversity of the generated synthetic images, assessing their impact on model performance. The analysis includes a comparison of generated images with real images to evaluate their effectiveness in addressing data insufficiency and enhancing feature extraction. In Section 5.3, the focus shifts to the explainability and regularisation aspects of the model. This section presents the results of various

manifold learning algorithms, evaluated using metrics such as Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). The discussion includes an examination of how these metrics contribute to model robustness and interpretability. Additionally, the explainable AI (XAI) module results are presented, demonstrating how explainability techniques have been integrated into the model to enhance transparency and understanding of the decision-making process. Finally, Section 5.4 provides a detailed presentation and discussion of the final classification results. This section includes an evaluation of the model's performance based on several metrics, such as sensitivity, specificity, accuracy, precision, recall, F1-score, and others. The results are compared with existing state-of-the-art methods to highlight the effectiveness of the proposed framework. The chapter concludes with a summary of key findings and their implications for colorectal polyp classification and future research directions.

Finally, Chapter 6 provides a comprehensive summary of the research conducted throughout this thesis, drawing well-supported conclusions based on the findings and analyses presented in previous chapters. Section 6.1 offers an in-depth exploration of the principal discoveries made during the research. This section synthesises the results from various experiments and methodologies, emphasising the significant contributions of the proposed framework to the field of colorectal polyp classification. It discusses how integrating synthetic image generation, advanced feature extraction, and explainable AI has addressed critical challenges, such as data insufficiency, feature representation, and model interpretability. The key findings highlight the effectiveness of the proposed approach in improving classification accuracy, enhancing model sensitivity, and providing a more transparent decision-making process for pathologists. Section 6.2 delves into the broader implications of the research, outlining how the results impact medical

image analysis and computer-aided diagnosis. This section discusses the potential benefits of the proposed framework for clinical practice, particularly in enhancing the efficiency and reliability of colorectal cancer screening and diagnosis. It also explores how the methodology can be adapted for other medical imaging tasks and domains, potentially leading to advancements in precision medicine and automated diagnostic tools. In Section 6.3, the thesis identifies and discusses the limitations of the proposed method. This section critically evaluates the challenges and constraints encountered during the research, including issues related to dataset quality, the computational demands of the model, and the generalizability of the findings. It also addresses areas where the methodology could be further improved or refined. Lastly, the chapter outlines promising directions for future research. It suggests potential avenues for enhancing the framework, such as exploring more advanced generative models, improving feature extraction techniques, and integrating additional explainability methods. The chapter concludes with recommendations for researchers and practitioners interested in building upon the work presented in this thesis, aiming to advance the field of colorectal cancer diagnosis and treatment.

1.7 Publications from Thesis

The following research papers have been written and published as a part of this research project.

Journal Publications

- Younas, F., Usman, M., & Yan, W. Q. (2023). An ensemble framework of deep neural networks for colorectal polyp classification. *Multimedia Tools and*

Applications, 82(12), 18925-18946. DOI: https://doi.org/10.1007/s11042-022-14177-0

- Younas, F., Usman, M., & Yan, W. Q. (2023). A deep ensemble learning method for colorectal polyp classification with optimised network parameters. *Applied Intelligence, 53(2)*, 2410-2433. DOI: https://doi.org/10.1007/s10489-022-03689-9

Chapter 2

Literature Review

In this chapter, we review four major themes of literature related to the contributions made in this thesis. First, we discuss previous research on colorectal polyp classification, data augmentation, and synthetic image generation. Second, we present the research on feature extraction in the medical domain. Third, we review the research on explainable AI in computer-aided diagnostic systems. Fourth, we review the existing feature selection methods and discuss the research work done on model regularisation and image classification. All of these themes form the basis for the research conducted in this thesis.

2.1 Colorectal Cancer Classification

Numerous models have been proposed for the computerized classification of colonic polyps, each aiming to improve the accuracy and reliability of detecting various polyps. One model was proposed to identify adenomatous and non-adenomatous polyp types, as outlined by Komeda et al. (2017). In this particular study, the data comprised 1,200 samples of adenomatous polyps and 600 samples of non-adenomatous polyps, totaling 1,800 samples. These samples were extracted from real-life colonoscopy examinations conducted at Kindai University Hospital. The dataset spans a period from January 2010 to December 2016, ensuring a comprehensive collection of cases for the model's development and testing. Computer-aided vision, a key component of this research, refers to a sophisticated tool used to classify objects within the visual field of medical imaging, specifically focusing on the identification and categorization of polyps. The integration of computer-aided vision with Convolutional Neural Networks (CNNs) in this work aims to enhance the accuracy of polyp classification significantly. The proposed CNN-based Computer-Aided Diagnosis (CAD) model demonstrated an accuracy of 0.75 on real-time endoscopy samples, which were rigorously tested using a 10-fold cross-validation method. However, it is important to note that this study does not delve into the classification of specific polyp types, such as serrated adenomas, adenomatous polyps, and hyperplastic polyps, which could be crucial for more precise medical diagnoses.

However, the authors suggest that model performance might have been further enhanced given access to a larger and more varied dataset. The problem of scarcity of data may partly be addressed by using methods like transfer learning, image augmentation or oversamplin,g which are meant to increase the capacity of the model as well as improve its generalisation ability to new data. Furthermore, fine-tuning the hyper-parameters of CNN

might help to achieve improved accuracy and overall model performance. Consequently, though the current accuracy of this model may not meet the highest standards, the CNN-CAD system still remains a preferable and promising tool for improving computerized polyp classification in medical imaging..

Patino-Barrientos et al. (2020) presented a deep learning (DL) model developed based on Kudo's classification system, which aims to identify polyps as either malignant or non-malignant. The dataset in this study comprised 600 images collected from 142 patients at Urudliz Hospital and the Biodonostia Health Research Institute. To address the issue of limited data, the researchers applied data augmentation techniques, thereby increasing the dataset size and diversity, which is crucial for enhancing the model's performance and robustness. The problem was approached iteratively, beginning with the implementation of a small, deep convolutional model to establish a baseline for performance. Subsequently, the outcomes were compared by implementing the VGG-16 neural network architecture, which is known for its depth and complexity in feature extraction. The researchers didn't stop there; they further fine-tuned the model to optimize its performance and compared the results across different configurations. The evaluation metrics employed in this study included accuracy, precision, recall, and F1 score, which together provide a comprehensive assessment of the model's classification capabilities. After fine-tuning, the model achieved an accuracy of 0.83, a precision of 0.81, a recall of 0.86, and an F1 score of 0.83, indicating a well-balanced performance across all metrics. Moreover, the performance of the proposed model was benchmarked against other traditional classification techniques, including Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). While the results obtained from SVM and KNN were comparable, the proposed deep learning model demonstrated a superior

approach to the classification task, particularly in terms of accuracy and generalisation to new data. However, despite these promising results, the study also highlighted the need for more data to further improve the model's performance. Utilising additional data, whether through further data augmentation or oversampling techniques, could significantly enhance the model's performance, leading to more reliable and accurate classification outcomes. The research emphasizes that deep learning models, while powerful, are heavily reliant on large and diverse datasets to achieve their full potential in medical image classification tasks.

Furthermore, an advanced artificially intelligent detection and classification model utilizing a deep neural network architecture has been developed by Ozawa et al. (2020) specifically for the classification of colorectal polyps. The algorithm underlying this method is based on a Single Shot Multibox Detector (SSD), which is a type of deep learning model designed for object detection. This model is particularly noteworthy because it takes into account various polyp classes, including adenoma, hyperplastic polyps, sessile serrated adenomas, and others, thereby contributing significantly to the classification of polyp subtypes. The study employed a substantial dataset for training and testing, comprising data from 12,895 patients who underwent colonoscopies at the Tada Tomohiro Institute of Gastroenterology and Proctology in Japan. A total of 16,418 images were used to train the CNN algorithm, including 3,021 images of polyps and 4,013 images of normal colorectal tissue. The processing time of the CNN was remarkably efficient, with an average of just 20 milliseconds per frame, highlighting the model's potential for real-time applications. The trained model demonstrated impressive performance by accurately identifying 1,246 polyps, achieving a sensitivity of 0.92 and a positive predictive value (PPV) of 0.86.

It is also worth noting that imaging modality slightly changed the model's performance. In terms of white-light (WL) images, both sensitivity and PPV were 0.90 and 0.83, respectively. The sensitivity and PPV values for Narrow Band Imaging (NBI) were even higher, i.e., 0.97 and 0.98, respectively. In general, 83% of polyps were correctly classified by the system, and WL imaging was able to identify as many as 97% of adenomas accurately. The study has emphasized this model's ability to detect and differentiate different kinds of polyps with great accuracy.

However, it is important to note that the study suggests the model's performance could be further enhanced. Another point noted by researchers was not all optimized hyperparameters have been used fully which show more possibility of making improved outcomes through refined tuning. Despite this, the results of the developed system are already impressive, demonstrating great potential for AI-based automated systems in the detection and classification of colorectal polyps. However this research shows an incredible development in being done with the help of such systems assisting medical professionals in diagnosing patients accurately besides fastening this process hence improving health care. It is suggested that further improvements need to be made on these areas through more fine-tuning since there still remains potential for getting better output from this model based on its training data features depictions. The instance given demonstrates how such models are useful when determining whether or not an individual has cancer, thus enhancing early diagnosis before it becomes too advanced, leading to severe conditions like death. Furthermore, if implemented widely, such a program could prove one essential solution towards addressing current challenges experienced globally across most healthcare institutions today. An earlier version indicating what we have now can be called artificial intelligence-aided colonoscopic image analysis

using deep learning-based algorithms. These results showed that up to four out of five human specialists achieved higher performance when assessed individually and as a team compared to the AI system.

A recently suggested model by Wei et al. (2020) leverages deep neural networks for classifying colorectal polyps, representing an advanced approach in medical imaging. The project employed a ResNet neural network architecture as the foundation for an ensemble model. This ensemble model was particularly sophisticated, comprising five distinct ResNets with varying depths: 18, 34, 50, 101, and 152 layers. Using multiple ResNets in an ensemble configuration allowed the model to capitalize on the strengths of different architectures, thereby enhancing its overall performance and robustness. The proposed model was designed to classify four critical types of colorectal polyps, which are commonly associated with varying degrees of malignancy risk. These types include tubular adenoma, tubulovillous or villous adenoma, hyperplastic polyp, and sessile serrated adenoma. Accurate classification of these polyp types is crucial for effective patient management and treatment planning, making developing such models a significant step forward in colorectal cancer prevention and diagnosis.

Dartmouth-Hitchcock Medical Center in Lebanon, New Hampshire, provided a diverse dataset collected by the researchers for training and testing the model. For this purpose, the dataset was carefully divided into subsets where 326 slides were used to train the model, 157 internal data slides were kept for further testing and validation was done using 25 slides. Also, they gathered an additional 238 slides from twenty-four different institutions so that their research could be exposed to a range of samples in order to avoid overfitting problems. This comprehensive approach to data collection and model training enabled the researchers to develop a highly effective deep learning model

capable of accurately classifying multiple types of colorectal polyps. The ensemble model used several ResNet architectures together with extensive and diversified datasets outlining the potential of deep learning in improving medical diagnostics. With advancing technology, such models are likely to become increasingly important tools for healthcare professionals, helping them to detect and classify colon polyps more precisely hence, improve patients' outcomes in the future.

A hierarchical algorithm was designed to classify colorectal polyps using DL models for medical image analysis as a structured approach. This was done by splitting the slides into smaller patches using a sliding window algorithm, which enabled the model to focus on local regions within the slides. In addition, these patches were classified by neural networks for detailed analyses of specific parts in each slide. These grid searches involve determining thresholds for every type of polyp, and they systematically test various parameters over ranges in order to ascertain optimal settings. The major aim of this study was to evaluate how well the hierarchical DL model performed compared to the results given by pathologists, who are the gold standard in medical diagnostics. The evaluation of the model's performance relied on key metrics such as accuracy, sensitivity and specificity, which are essential indicators of how effective these models could be in real clinical applications. The outcomes showed promise, where on average, 0.93 accuracy was found within our internal dataset for the model slightly outperforming pathologists' 0.91 accuracy rate achieved respectively. The model was 87% accurate for the external dataset, which is equivalent to that of pathologists at 86%. This means that it is the same as a well-trained medical doctor and can therefore be used in healthcare as an important tool.

However, this study had one major limitation: the dataset size was relatively

small, and thus, its performance in real-world scenarios would be affected where there is more variability and complexity. It is widely known that acquiring enough medical data is a big challenge; these authors propose transfer learning as one solution to such thorny issues as oversampling. In situations with few data sets available, such as in healthcare systems, transfer learning may be useful. Additionally, it is recommended that one use artificial techniques such as data augmentation to have more examples and increase the diversity of the dataset, hence making it robust. Finally, the hierarchical DL model results agreed with those of local pathologists, proving its potential as an assistive tool for doctors involved in the evaluation and classification of colorectal polyps. Through its accuracy and reliability in classification, this model could help the diagnostic process to be straightened out, leading to better patient outcomes and the more efficient consumption of healthcare services.

The approach developed by Korbar et al. (2017) leverages deep learning, a sophisticated computer-guided analysis method, for the classification of CRC polyps. This model is designed to identify five specific types of CRC polyps: hyperplastic polyps, sessile serrated adenomas, traditional serrated adenomas, tubular adenomas, and tubulovillous/villous adenomas. The dataset used in this study was substantial, comprising 458 whole-slide images for training and 239 images for testing, all of which were obtained from patients diagnosed with colorectal cancer. In total, 2,074 cropped images were collected and used to train and validate the model, ensuring a diverse and comprehensive dataset for analysis. To determine the most effective approach for CRC polyp classification, the researchers executed and evaluated various deep-learning neural network architectures. Standard network architectures such as AlexNet, VGG, GoogleNet, and ResNet were all tested to characterize the polyps. Among these, ResNet

stood out for its superior performance, particularly when implemented with varying numbers of computational layers. Specifically, the ResNet architecture was deployed in multiple configurations: ResNet-A, B, C, and D, which comprised 50, 101, 152, and 152 layers, respectively. Although ResNet-C and D both 152 layers, they differed in their approach to mapping, with ResNet-C employing identity mapping and ResNet-D using projection mapping.

Out of all the tested architectures, ResNet has shown the highest accuracy in classifying the CRC polyps. The model was evaluated using multiple validation metrics like accuracy, precision, recall, and F1-score. This yielded impressive results since the model had an accuracy of 0.93, a precision of 0.89, a recall of 0.88 and an F1-score of 0.88. These figures imply that the model was very successful in accurately recognizing and classifying different types of polyps; hence, it is a valuable tool for diagnosing and treating colorectal cancer patients. Nevertheless, what is indicated by the research is that it would have been better if there were more hyperparameter tuning performed on this model, and the study revealed that there were several areas where better hyperparameter tuning could improve performance by fine-tuning hyperparameters. Therefore, while they are important to consider while developing a good architecture, we need to consider both their selection and setting up to get the best result possible. This nonetheless shows that despite deep learning offering potential for various improvements in medical diagnoses Korbar et al. (2017), this approach still represents a great leap forward in utilizing deep learning techniques for CRC polyp classification, thereby improving medical diagnostics accuracy through enhanced technology deployment efficiency.

Another approach suggested by Mesejo et al. (2016) is to facilitate clinical workflows with a virtual biopsy technique by distinguishing colorectal polyps. This novel

method combines ML and computer vision (CV) algorithms for categorizing hyperplastic, serrated, and adenoma polyps. The process begins with inputting images in which key features are extracted for further analysis. The next step is the use of the structure of motion methods to rebuild 3D lesions and extract 3D shape characteristics that add spatial dimensionality to the analysis, improving classification performance. Once 3D features have been extracted, the images are supplied to the system for classification purposes. Thus, these classifiers were tested against another ensemble classifier called Support Vector Machines (SVM), a widely used machine learning algorithm known for good performance in classification tasks. For training and testing their models, the researchers compiled a dataset consisting of 76 colonoscopy clips. The various classification methods demonstrated different levels of accuracy in the results. On average, the Random Forest classifier outperformed the other methods with an accuracy rate of 0.78. Although it scored lower than its counterpart at 0.29, Random Subspaces made some improvements by achieving an accuracy rate of 0.49 while SVM had an accuracy rate of 0.29. The practical applicability of models was evaluated, and these results were compared with the performance of expert and beginner practitioners. This study highlights possible applications for virtual biopsy in classifying colorectal polyps that could lead to improving clinical workflows through increased precision and efficiency rates. However, some classifiers, such as SVM, have shown relatively poor performance, which necessitates further adjustment and optimization on this front. Nonetheless, integrating ML and CV into this technique is a significant advancement towards technology-assisted medical diagnostics, as well as offering a potentially useful addition to medical professionals' expertise in formulating better diagnosis tools, despite low performances by certain classifiers such as SVMs indicating areas requiring additional model refinement and optimization efforts on this matter.

Moreover, another computer-aided technique proposed by Zhang et al. (2016) focuses on the identification and categorization of hyperplastic and adenomatous colorectal polyps, utilizing Convolutional Neural Networks (CNNs) for both detection and classification tasks. This approach employs CNNs in a two-step process, starting with the detection of polyps and subsequently classifying them into specific categories. Unlike the method described in the study by Komeda et al. (2017), which also used CNNs for polyp classification, the approach taken by Zhang et al. (2016) differs in how the classification problem is addressed. Initially, a CNN model is implemented to detect polyps in colonoscopy images. Once the polyps are detected, a second CNN model is employed to classify the detected polyps into either hyperplastic or adenomatous types, which are crucial for determining the potential malignancy of the polyps. The features necessary for this classification task are derived from two extensive public datasets: the ILSVRC (ImageNet Large Scale Visual Recognition Challenge), which contains 1.2 million images, and Place205, which includes 2.5 million images. These large datasets provide a rich source of diverse features that enhance the model's ability to accurately classify colorectal polyps. It is important to note that this approach yields results worth attention, and the system attains an accuracy of 0.85 compared to practitioners' rate of 0.74. This step justifies the fact that in some cases, computer-assisted approaches can outperform humans in certain diagnostic tasks, such as early polyp diagnosis, where accurate detection and classification are vital. However, the study also suggests that further optimization of hyperparameters might have achieved better performance, indicating possibilities for model improvement. Nevertheless, it is commendable that, unlike other techniques which necessitate more extensive data preparation, the proposed method has a minimal preprocessing requirement for the diagnosis of colorectal cancer. Thus, narrowing down the diagnostic process, making doctors use less time and improving

the quality of decisions about colorectal cancer diagnoses earlier and accurately with these computerized aids will ensure prompt treatment for patients, hence a chance to live longer.

For classifying colorectal polyps into two main types (adenomatous and serrated), a deep learning model was presented by Zachariah et al. (2020). This study aimed to achieve low cost, timely polyp classification, while at the same time offering doctors with a tool that augments their diagnostic acumen. The intention is to streamline the process of classification through improved clinical workflows so as to ensure that more accurate and timelier medical decisions are made. The model has been trained and tested on an extensive dataset consisting of 5,278 high-quality images which provided a solid foundation for developing a robust classification system. This proposed CNN model architecture can be divided into two core components: the base module and the head module. The Inception-ResNet-v2 architecture is applied in designing the base module because it facilitates the extraction of complex details from images due to its deep and complicated structure. These features are, in turn, passed onto the head module, where they are transformed into grades on an intensity scale. After pattern recognition, these grades enable differentiation between adenomatous or serrated polyps. This investigation is the comparison of two different imaging methods, including white light imaging (WL) and narrow-band imaging (NBI). Although the techniques used in these two methodologies are different, no difference was observed between WL and NBI in terms of the model's performance. Therefore, it can be concluded that this model can be used with any type of imaging modality that is needed in clinical practice. The dataset showed a high NPV of 97%, which means that the model can be trusted to correctly exclude polyps from non-defined groups. Also, the overall agreement on surveillance was

94%, which showed good potential for this model to become a reliable tool for assisting in colorectal polyp monitoring and classification. However, further studies could benefit from including other types of polyp classification. This expansion would help increase its usefulness in clinical settings since it could identify and categorize many more types of polyps, leading, therefore to the provision of better diagnostic results possible through a complete diagnosis process.

Another significant contribution to the classification of colorectal polyps was developed by Poudel et al. (2020), targeting five specific classes: Adenocarcinoma, Adenoma, Crohn's disease, Ulcerative colitis, and normal images. This study involved collecting data from Gill Hospital, which yielded a dataset of 3,515 image samples. To further validate the proposed model, the researchers also used the KVASIR dataset, which comprises 4,000 images. The inclusion of this diverse and extensive dataset helped ensure the robustness and generalisability of the model in different clinical scenarios. The deep learning model designed in this study employs deep layers that preserve spatial information through the use of dilated convolution. This technique enhances the model's ability to classify polyps more accurately by capturing detailed spatial features, which are crucial for distinguishing between different types of colorectal conditions. To avoid overfitting, which can be a common issue in deep learning models, the researchers utilised ResNet-50 as the backbone of the model. Additionally, Drop Block, a regularisation technique, was employed to improve the model's generalisation capabilities by randomly dropping entire blocks of feature maps during training, forcing the model to learn more robust features. The model's performance was assessed based on some key metrics such as accuracy, recall, precision, and F1-score. The results were encouraging with the colorectal dataset having an F1 score of 0.93 and KVASIR having a score of 0.88.

These high F1 scores point to a successful balance between precision and recall by the model, thereby accurately classifying colorectal polyps and related conditions. However, despite the strong performance of the proposed method, future research would benefit from contrasting it with other deep learning architectures. These comparisons may reveal more about the strengths and weaknesses inherent in this model to obtain even better classification results. Researchers can decide which is most effective for specific tasks within colorectal polyp classification, thus adding to the accurate diagnostic tools expedient in clinical practice by exploring different architectures.

A transfer learning architecture was proposed to enhance the classification of colorectal polyps, leveraging advancements in deep learning techniques. The dataset for this study consisted of 1,000 instances collected from patients during colonoscopies at Gachon University Gil Hospital. To evaluate the effectiveness of the proposed method, the authors compared it with AlexNet using various configurations, including AlexNet alone, AlexNet + SOS, AlexNet + ImageNet, AlexNet + Places, and their own proposed method, NIN + ImageNet. The term "Network in Network" (NIN) refers to a deep learning architecture that incorporates multiple fully connected layers or multilayer perceptrons within convolutional networks. This approach allows for more complex feature extraction compared to traditional CNNs, leading to improved performance in classification tasks. The proposed NIN + ImageNet model demonstrated a significant accuracy enhancement of 18.9% over the models based on AlexNet, highlighting the effectiveness of this architecture in improving classification results. A recall rate of 0.92 was achieved by the proposed model, with a standard deviation of 0.029 and an area under the curve (AUC) close to 0.930 with a standard deviation of 0.020.According to these measures, it is evident that the model performs well in discerning between normal and abnormal polyps,

hence assisting pathologists in proper diagnosis.The research however proposes that comparing this model to other advanced architectures like ResNet and DenseNet can offer additional insights on how it performs relatively.This would therefore provide further proof for the effectiveness of NIN-based approach as well as uncover more areas where it could be improved.In addition, inclusion of classification for more types of polyps would enhance its clinical applicability and practicality.Moreover, classifying a wider range of polyp types could potentially increase its usefulness in clinical contexts.Finally, the proposed architecture for transfer learning demonstrates great potential towards enhancing accuracy and efficiency in colorectal polyp classification.

A method using a stacking ensemble approach for improved polyp classification performance was proposed by Rahman et al. (2021). The dataset used in this study was sourced from the University of Alcala and consisted of 26,512 images categorized into four classes: Hyperplastic, Serrated, Adenoma, and Non-polyp. Preprocessing steps included removing glare from images, which, while typically beneficial, was noted to potentially negatively impact classification performance in this case. To address this, a frame selection method was implemented to reduce the model's processing time, optimizing the efficiency of the classification process. The core of the proposed method involved a stacked ensemble approach that combined three convolutional neural network (CNN) architectures: Xception, ResNet-101, and VGG-19. Each of these architectures was optimized individually before being integrated into the ensemble. A softmax classifier was applied to predict the likely outcome of each model, leveraging the combined strength of the different architectures. The ensemble neural network performed best with two hidden layers, comprising 10 and 8 neurons respectively, and the ReLU activation function. The performance metrics achieved were impressive: accuracy of

98.5% ± 0.2%, recall of 96.2% ± 0.87%, precision of 92.1% ± 4.62%, specificity of 98.9% ± 0.36%, and an area under the curve (AUC) of 0.9912. These metrics demonstrate that the stacked ensemble approach significantly outperformed individual neural network models. Despite the strong performance of the proposed method, the authors noted that comparing it with other advanced architectures could provide further insights and potentially reveal additional improvements. This comparison could help validate the effectiveness of the stacking ensemble approach and identify opportunities for further enhancement.

The study conducted by Sakr et al. (2022) proposes a novel and efficient deep learning approach based on a Convolutional Neural Network (CNN) for automatic early detection of colon cancer from histopathological images. The main objective of the study is to propose a novel and efficient deep learning approach for automatic early detection of colon cancer from histopathological images. The study uses a Convolutional Neural Network (CNN) model to detect colon cancer from histopathological images, with the input images being rescaled and normalized before feeding them into the model. The study employs seven deep architectures with different batches and layers and selects the best performance architecture as the proposed model. The proposed model consists of four convolution blocks, a flatten layer, a fully connected layer, and a SoftMax layer, using the Adam optimizer and Cross-entropy loss function. The proposed approach achieves a high accuracy of 99.50% in detecting colon cancer from histopathological images, with the results being compared to existing state-of-the-art methods. The proposed model achieves high accuracy, precision, recall, and F1-score of 99.50%, 99%, 100%, and 99.49%, respectively. The model can detect all abnormal images correctly with an accuracy of 100% and accurately detect 99% of the normal images. The study concludes

that the proposed deep learning approach is efficient and effective in detecting colon cancer from histopathological images, with a high accuracy of 99.50%. The proposed method is a new efficient deep CNN approach for colon cancer detection, outperforming most previous deep learning approaches. The model is lightweight and can be useful in cases where pathologists need to be insured. However, the study did not address the challenges collectively.

The diagnosis and treatment of colon cancer are significant social and economic challenges due to high mortality rates, with various tests and imaging modalities available for screening and diagnosis, including histopathology images, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). The main goal of the study done by Tharwat et al. (2022) is the automatic diagnosis of colon cancer with high detection accuracy and without manual intervention. The study aims to provide a comprehensive review of the current studies on colon cancer, classified into deep-learning (DL) and machine-learning (ML) techniques. The study uses various imaging modalities, including Computed Tomography (CT), Endorectal Ultrasound (ERUS), virtual Computed Tomography Colonoscopy (CTC), and Magnetic Resonance Imaging (MRI). The study also uses machine learning (ML) and deep learning (DL) techniques to analyze the images. The study various ML and DL techniques, including CNNs, SSD-GPNet, and SVM, across different datasets and imaging modalities to evaluate their effectiveness in cancer detection. The study found that various machine learning (ML) and deep learning (DL) techniques can be used to accurately detect colon cancer. The study also found that the use of a combination of CT with PET produces better results in oncology to diagnose and detect the stage of cancer. The results demonstrated high accuracy and sensitivity in detecting colorectal cancer, with some methods achieving up to 99.10% accuracy and

98.82% sensitivity. The study concludes that the diagnosis and treatment of colon cancer are significant social and economic challenges due to high mortality rates. The study also concludes that various machine learning (ML) and deep learning (DL) techniques can be used to accurately detect colon cancer. The study concludes that ML and DL techniques are effective in early-stage cancer detection, which can lead to early treatment and reduced mortality rates. However, they did not consider all the major changes of biomedical in a single pipeline.

Khazaee Fadafen & Rezaee (2023) explores the use of a multi-SVM with an RBF-based learning pool in an ensemble learning framework to improve the classification accuracy of colorectal cancer (CRC) histopathological images (HIs) using deep learning techniques like dResNet and ResNet architectures. The objective of the study was to develop a hybrid architecture for classifying colorectal cancer based on histopathological images, which can perform effectively in the absence of annotated datasets and minimize pathologists' workload. The study also aimed to reduce computational complexity and improve the generalisability of the model. A dilated ResNet structure and attention module were used to generate deep feature maps, and neighborhood component analysis (NCA) was employed to overcome computational complexity. Data was fed into a deep support vector machine (SVM) based on an ensemble learning algorithm called DeepSVM. The study employed various deep learning architectures, including CNNs, ResNet, DenseNet, and VGG, along with techniques like dilated convolution, attention mechanisms, and ensemble methods to enhance feature extraction and classification accuracy.

The study used an ensemble learning framework with multi-SVM classifiers and RBF kernels. The datasets were divided into training, testing, and validation sets

in a 60–20–20 ratio. The dResNet-101 model was used for feature extraction, and the classification process was performed using 50 features. The study used a dilated ResNet-101 model with an attention module for automatic feature generation and learning, and a NCA-based feature selection procedure to reduce computational complexity. The selected features were then fed to a robust DeepSVM classifier to classify colorectal cancer multi-class texture. The hybrid model achieved 98.75% and 99.76% accuracy on CRC datasets, demonstrating its effectiveness in classifying colorectal cancer tissues. The proposed methods achieved high classification accuracy, with some models reaching up to 99.75% accuracy. The use of dilated convolution and attention mechanisms improved feature fusion and spatial resolution, leading to better performance in detecting CRC.

Alboaneen et al. (2023) conducted a review of the state-of-the-art research on AI-based ML and DL techniques applied to the modelling of colorectal cancer, highlighting their potential in early diagnosis and treatment. The objectives of the studies include developing predictive models using ML or DL approaches to predict a normal or abnormal state, either using a public or collected dataset. The studies also aim to investigate the application of ML and DL algorithms in colorectal cancer detection and diagnosis. The methods used in the study include a systematic review of 57 research articles on AI-based ML and DL techniques for colorectal cancer diagnosis. The methods used in the studies include the application of ML and DL algorithms, such as convolutional neural networks (CNNs), support vector machines (SVMs), and random forests (RFs). The studies also used various datasets, including public datasets, and collected datasets. The results of the study show that ML and DL models can achieve high accuracy rates in predicting colorectal cancer, with some models outperforming others. The results of the studies show that ML and DL algorithms can achieve high accuracy in predicting colorectal

cancer, with some studies achieving accuracy rates of over 99%. The studies also found that the type of dataset used can impact the accuracy of the predictions. The conclusions of the study are that AI-based ML and DL techniques have the potential to improve the diagnosis and treatment of colorectal cancer, and that further research is needed to develop more accurate and reliable models. The conclusions of the studies include the high accuracy of ML and DL algorithms in predicting colorectal cancer, and the potential of these algorithms to improve the prediction of malignancies like colorectal cancer. The studies also highlight the need for further research in the field of ML and DL applications in colorectal cancer prediction.

The research conducted by Sobur & Rana (2024) focuses on advancing cancer classification using a hybrid deep learning model for lung and colon cancer detection. The study is a dataset of 20,000 high-resolution histopathological images across five classes. The hybrid model incorporates dual Convolutional Neural Networks (CNNs) for feature extraction, followed by innovative feature reduction techniques to enhance computational efficiency and robustness. Integrating global average pooling and dense layers with dropout regularisation prevents overfitting and ensures generalisability. The model achieved an impressive 99% classification accuracy, showcasing its precision in handling high-dimensional datasets. The study underscores the importance of deep learning in medical image analysis, particularly in cancer classification. It addresses the challenges posed by high-dimensional datasets in histopathological images and highlights the significance of innovative model architecture and data processing techniques. Feature reduction methods like Principal Component Analysis (PCA) play a vital role in reducing dimensionality, enhancing model efficiency, and improving classification accuracy. The research emphasizes the essential role of advanced computational models in medical

imaging, especially in early cancer detection and effective treatment planning. Overall, the paper presents a comprehensive overview of the advanced hybrid deep learning model's design and evaluation for lung and colon cancer classification. The model's exceptional performance metrics and robustness underscore its effectiveness in accurately distinguishing between different tissue types and cancer conditions. The authors suggest potential areas for future optimization, including exploring additional feature reduction techniques, optimizing model architecture, and investigating error cases to further enhance the model's capabilities in medical image analysis. Investigation into additional feature reduction techniques: Explore and compare different dimensionality reduction methods beyond PCA and autoencoders to further enhance the model's performance in handling high-dimensional histopathological images.

A recent study done by Vanitha et al. (2024) focuses on enhancing diagnostic imaging of lung and colon cancers through a novel deep learning framework that combines the Xception and MobileNet architectures. The research addresses challenges such as accuracy and overfitting in the classification of these cancers. The proposed ensemble model aims to improve feature extraction, model robustness, and reduce overfitting by training on a comprehensive dataset of histopathological images and validating against a balanced test set. Results show an impressive classification accuracy of 99.44%, with perfect precision and recall in identifying cancerous and non-cancerous tissues, outperforming traditional methods. The integration of Gradient-weighted Class Activation Mapping (Grad-CAM) enhances interpretability, allowing clinicians to visualize the diagnostic reasoning process. Through visual explanations, the model's decision-making process becomes transparent, crucial for clinical acceptance and personalized treatment planning. The study not only expands medical imaging technology but also sets the

stage for further research in cancer diagnostics. The research emphasizes interpretability through Grad-CAM, enabling clinicians to verify model predictions visually. Despite the model's strong performance, potential limitations related to dataset biases and generalisability to diverse clinical environments are noted. The study suggests future applications in other cancer types and medical imaging tasks, aiming to advance automated and reliable diagnostic processes for improved patient outcomes in oncology. The review of detection and classification techniques for colorectal polyp classification are summarized in Table 2.1. This review provides a comprehensive overview of the current state of methods and technologies in this field.

2.2 Synthetic Image Generation

Generative Adversarial Networks (GANs) represent a cutting-edge approach in training generative models. GANs consist of two key components: the Generator (G) and the Discriminator (D). The Generator (G) learns to capture and reproduce the data distribution, effectively generating new samples that resemble the training data. Conversely, the Discriminator (D) aims to distinguish between samples from the real training dataset and those generated by G. This adversarial process helps G improve its ability to generate realistic samples over time. A notable variant of GANs used in medical image analysis is the Conditional GAN (cGAN). cGAN extends the standard GAN framework by incorporating additional information from a specific class or condition during training. This approach allows both the Generator and the Discriminator to utilise conditional data, which enhances the model's ability to generate more targeted and relevant samples.

One foundational work in this area was developed by Mirza & Osindero (2014),

Table 2.1: Review of detection and classification techniques of CRC diagnosis

Author(s)	Focused Area	No of images	Polyp Types	Data Source	Network	Evaluation results
Komeda et al. (2017)	Classification	1800	Adenomatous Non-adenomatous	Kindai University Hospital	CNN	Accuracy: 0.75
Patino-Barrientos et al. (2020)	Classification	600	Malignant Non-malignant	Biodonostia Health Research Institute	VGG-16	Accuracy: 0.83 Precision: 0.81 Recall: 0.86 F1 score: 0.83
Ozawa et al. (2020)	Detection Classification	16,418	Adenoma Hyperplastic Sessile serrated	Tada Tomohiro Institute of Gastroenterology and Proctology, Japan	CNN	Sensitivity: 90 PPV: 83
Wei et al. (2020)	Classification	508	Tubular Adenoma Tubulovillous Hyperplastic Sessile serrated Adenoma	Dartmouth-Hitchcock Medical Centre, Lebanon, New Hampshire.	Ensemble of ResNets 18, 34, 50, 101, 152 layers	Accuracy: 0.87
Korbar et al. (2017)	Classification	2,074	Hyperplastic Sessile serrated Traditional serrated Tubular Tubulovillous	N/A	ResNet50	Accuracy: 0.93 Precision: 0.89 Recall: 0.88 F1-score:0.88
Mesejo et al. (2016)	Classification	76	Hyperplastic Sessile Serrated Adenoma	UCI Repository	SVM Random Forest Random subspace	Accuracy: 0.78 0.49 0.29
Zhang et al. (2016)	Detection Classification	1.2 million 2.5 million	Hyperplastic Adenomatous	ILSVRC	CNN	Accuracy: 0.85
Zachariah et al. (2020)	Classification	5278	Adenomatous Serrated	N/A	Inception-ResNetv2	Negative predicted value: 0.97
Poudel et al. (2020)	Classification	35154000	Adenocarcinoma Adenoma Crohn's Disease Ulcerative Colitis Colitis	Gill Hospital KVASIR dataset	ResNet50	F1-score: 0.93 0.88
Rahman et al. (2021)	Classification	26,512	Hyperplastic Serrated Adenoma Non-Polyp	University of Alcala	Xception Ensemble ResNet-101 VGG-19	Accuracy: 0.98 Recall: 0.96 Precision: 0.92 Specificity: 0.98

who introduced the concept of cGANs. Their model was designed to estimate image distributions conditioned on auxiliary information. For example, in their research, they applied this approach to the MNIST digits dataset, where the additional class labels served as the conditioning information. This allowed the Generator to produce images corresponding to specific digits, improving the relevance and quality of generated samples. While the cGAN approach represents a significant advancement, adding an initial classifier to focus on diversity and data quality could further enhance the effectiveness of the model. By incorporating a classifier that evaluates the diversity and quality of generated samples, researchers could ensure that the dataset not only matches the desired conditions but also maintains a high level of variation and fidelity. This additional step would contribute to the creation of a more robust and high-quality dataset, further advancing the capabilities of generative models in various applications, including medical image analysis.

To generate realistic images, class-conditional GANs (cGANs) such as BigGAN have shown remarkable ability by exploiting class-specific information. Notably, BigGAN stands out as it enables one to produce high-resolution pictures with fine details and realism that are important tools in various domains such as medical image analysis. However, one of the major concerns when working with cGANs is the fact that it requires substantial amounts of good quality training data. This may result in poor performance if there is inadequate training data where the model may generate lower quality or less diverse images. In medical image analysis, this problem can be significant because annotated datasets are often limited and can determine how good the images produced will be. DiffAugment presents a way to improve and stabilize the process of training cGANs according to S. Zhao et al. (2020). To make GANs more robust, DiffAugment

uses differentiable augmentations so that generated images become more diverse. By doing this, issues related to scarcity of data are alleviated and overall model performance is improved through training data augmentation compatible with the GAN framework.

Even with the advanced cGAN models, the problem of producing low-quality images persists. This emphasizes the necessity of collecting images of various categories and of superior quality. It is essential to have all necessary training data available so as to enable any clinical conditions and anomalies of the medical imaging to be appropriately modeled. Having high-quality datasets helps to improve the quality of image synthesis, which in turn improves the efficiency of GANs as well as other generative models for use in the classification of medical images. To conclude, although in the recent years advanced cGANs have been able to generate life-like images with relative ease, it remains apparent that continuous endeavors toward data enhancement and the refinement of training techniques are needed when tackling medical image classification amongst other tasks.

A framework for classifying colorectal cancer (CRC) histopathological slides using a generative adversarial network (GAN) within a semi-supervised learning approach was proposed by Sasmal et al. (2023). This innovative framework leverages both labelled and unlabelled data to enhance classification performance. In this approach, the GAN's discriminator is trained to perform two distinct roles: Supervised settings, where, the discriminator classifies histopathological images based on the labelled data. This helps the model learn specific features related to CRC classification and unsupervised settings where, the discriminator distinguishes between real and synthetic images, similar to traditional GANs. This training allows the discriminator to extract nuanced features from unlabeled images, contributing to the development of a more versatile and robust

model. By combining supervised and unsupervised learning, the framework benefits from the strengths of both approaches. The supervised learning component ensures accurate classification using labelled data, while the unsupervised learning component enhances feature extraction and generalisation by leveraging the extensive unlabeled dataset.

In a related effort, MobileNetV2 was for effective CRC classification (Raju & Rao, 2022). MobileNetV2 is a lightweight and efficient deep learning architecture that is well-suited for mobile and embedded applications. To address data imbalance challenges, a data augmentation technique based on CycleGAN was applied before classification. CycleGAN is effective in generating additional synthetic images that help balance the dataset, improving the overall performance of the classification model. Together, these approaches illustrate the potential of combining advanced GAN frameworks and deep learning architectures with innovative data augmentation techniques to enhance CRC classification and tackle common challenges in medical image analysis.

Yoon et al. (2022) proposed an advanced system for detecting polyps that focuses on capturing the morphological characteristics of sessile serrated lesions (SSLs). This system aims to improve the identification of overlooked or missed lesions by generating high-resolution samples of SSLs using Generative Adversarial Networks (GANs). The use of GANs helps address the issue of imbalanced polyp data by creating synthetic, high-quality images of SSLs, which enhances the training dataset and improves detection performance. Additionally, Chen et al. (2021) demonstrated how classifiers can be integrated with conditional Generative Adversarial Networks (cGANs) within a unified framework to enhance performance. Their approach involves decomposing the joint probability distribution to link the objectives of cGANs with classification

tasks. By incorporating a classic energy model to parameterize these distributions, the framework provides a systematic approach to combining classifiers with cGANs. This integration allows for a more comprehensive model that benefits from both generative and discriminative learning, ultimately improving the accuracy and effectiveness of the classification system.

Yu & Zhang (2020) proposed a novel model that incorporates a deep Conditional Generative Adversarial Network (cGAN) architecture, with a significant enhancement in the form of a capsule network used in place of the conventional Convolutional Neural Network (CNN) in the GAN's discriminator. This modification aims to address challenges related to variations in object pose and spatial relationships within the images. Capsule networks are designed to capture and preserve spatial hierarchies and relationships between features in a way that traditional CNNs might not. By utilizing capsule networks in the discriminator, the model benefits from improved handling of variations in input data, such as changes in object orientation and spatial configuration. This allows the GAN to generate and recognize images with greater accuracy and robustness, especially in scenarios where traditional CNNs might struggle with these variations. The integration of capsule networks into the GAN framework enhances the model's ability to learn more invariant image representations, leading to better performance in tasks such as image generation and recognition. This approach represents a significant advancement in leveraging GANs for complex image processing tasks by addressing limitations related to spatial relationships and pose variations.

Rau et al. (2019) employed the pix2pix Conditional Generative Adversarial Network (cGAN) to transform monocular endoscopic images into depth maps. This transformation is particularly useful for applications such as navigation and assessing

lesion sizes during colonic examinations. By converting 2D endoscopic images into 3D depth information, the system provides a more comprehensive view that aids in better navigation and measurement within the colon. To address the challenge of limited labelled training data in endoscopy, the authors proposed utilizing simulation environments. These environments generate synthetic data that can supplement the training process, helping to overcome data shortages. Additionally, they suggested training both the generator and discriminator components of the network using unlabeled video frames. This approach facilitates the adaptation of the model to real-world colonoscopy scenarios by leveraging available unlabeled data, thus improving the model's performance and robustness in practical applications. This method highlights the potential of integrating cGANs with depth estimation techniques and simulation environments to enhance endoscopic procedures and address common data limitations in medical imaging.

Shin et al. (2018) proposed a method that employs a combined input-conditioned image approach for training networks, utilizing edge filtering techniques to generate realistic polyp images while preserving the inherent structures of endoscopic frames. This technique produces synthetic samples from standard colonoscopy images, which are relatively easy to obtain, addressing the challenge of data scarcity. The generator network's encoding phase uses multiple dilated convolutions to capture large receptive fields and prevent feature map size contractions, maintaining detailed spatial information. During the decoding phase, convolutional layers are employed for upsampling, which minimizes artifacts in the generated images. This approach effectively combines realistic image generation with the preservation of endoscopic frame details, enhancing the quality and applicability of synthetic data for polyp detection and classification.

2.3 Feature Extraction

In the 1980s, Convolutional Neural Networks (CNNs) were first introduced, laying the groundwork for what would become a groundbreaking approach in the field of machine learning. Despite their early inception, CNNs gained substantial popularity and widespread use around 2012, particularly for their remarkable applications in image detection and classification, as well as in natural language processing (Ullman, 2000). CNNs are specifically designed to extract meaningful features from images while preserving their 2D spatial structure. This is achieved through the use of sliding filters, or convolutional kernels, that traverse the input data to detect various patterns and features. Over time, several methods based on CNNs have been developed to automate the classification of colon polyps, a critical task in medical imaging. For example, a model proposed by Komeda et al. (2017) focuses on categorizing polyps into two primary types: adenomatous and non-adenomatous. This model utilises computer vision techniques to enhance both object classification and polyp detection.

Integration of CNNs with computer-aided detection (CAD) systems, the model processes real-time images captured during colonoscopy procedures. Although this approach achieved a 75.1% accuracy rate in a 10-fold cross-validation test, the performance of the model could benefit from improvements through the use of a larger dataset. To address the challenge of data insufficiency, methods such as transfer learning and image augmentation can be employed. Additionally, fine-tuning hyperparameters can further enhance the model's performance and classification accuracy, leading to more reliable outcomes in the automated detection and classification of colon polyps.

Building on the success of Transformers in automated translation, convolution-free models relying solely on transformer layers have garnered significant attention in the computer vision domain (Vaswani et al., 2017). Originally developed for natural language processing (NLP), Transformers consist of encoder and decoder blocks, each comprising multiple layers of similar architectures. The encoder section of the transformer is responsible for encoding the input data, while the decoder generates the output sequence by leveraging the contextual information provided by the encoder. Each transformer block features a feed-forward neural network (FFN), a shortcut connection, and layer normalization (LN), which collectively enhance the model's ability to capture complex patterns and relationships within the data. In the realm of computer vision, the ViT represents a groundbreaking architecture designed to apply transformer principles to image classification tasks (Dosovitskiy et al., 2020). Unlike traditional CNNs, which rely on convolutional layers to extract features from images, ViT operates as a pure transformer model that processes sequences of image patches. In practice, ViT begins by transforming images into a sequence of patch tokens, which are then processed by the transformer model. A classification token (CLS), similar to the approach used in BERT for NLP (Devlin et al., 2018), is added to the sequence to facilitate the classification task. This innovative approach aims to match or even surpass the performance of CNNs in image classification, demonstrating the potential of transformer-based models in visual recognition tasks (Rangarajan et al., 2018).

A notable distinction between CNNs and ViTs lies in their final embedding processes. In CNNs, the final embedding is typically achieved by averaging the features across spatial locations, which helps in summarizing the spatial information into a fixed-size representation. In contrast, Vision Transformers utilises a different approach,

employing the classification token (CLS). This CLS token interacts with the sequence of patch tokens at each layer of the transformer encoder, allowing the model to capture and integrate information from various patches of the image (Borhani et al., 2022).

Although there are studies that have employed CNN models in colorectal polyp classification, it is possible that Vision Transformers may be a better approach. ViTs have also demonstrated great promise in tasks that include detection and segmentation of colorectal polyps as they are well suited to modeling intricate structures and relations in image data. However, not much has been done towards applying Vision Transformers in the classification of colorectal polyp images. Considering the convincing results obtained by transformers in the field of image classification, one would reason that this particular area warrants further exploration. By filling this void, the quality of automated classification of colorectal polyps may improve enabling clinicians to better control the condition.

An interesting method for finger image feature extraction is presented in Tandon & Namboodiri (2022) which is based on ”unified end to end transformer body”. This method is an improvement which makes it possible to carry out the extraction of overall characteristics, minutiae, and local and micro-features from samples of a single fingerprint one after the other. The main benefit of this approach is that there is no longer any need for employing multiple independent modules for each distinct feature extraction task. By putting a combined model containing all the functions required to perform the feature extraction task from the input images, which makes along with improvement the workflow of feature extraction tasks. The researchers demonstrated the effectiveness of this unified method through an efficient feature-matching process that leverages both global and local perspectives. This integration results in superior performance on various

datasets, setting a new standard in the field. This approach is particularly noteworthy as it showcases how combining different types of feature extraction within one framework can lead to more comprehensive and accurate results.

Our work is similar in its objective, as we also focus on advanced feature extraction techniques using a vision transformer model. By adopting a similar unified approach, we aim to leverage the strengths of transformer models for extracting a wide range of features in a single, cohesive process. This method not only simplifies the feature extraction pipeline but also improves the model's ability to capture and utilise div and complex features effectively, aligning with the successful strategies demonstrated in fingerprint analysis.

A dual-stream feature extraction network, known as DSFENet, has been proposed to enhance the accuracy of building extraction tasks (Xia et al., 2023). This network represents a significant advancement in the field by leveraging both CNN-based and transformer-based methods for semantic segmentation. The integration of these two approaches allows DSFENet to effectively capture both local and global features, thereby improving the accuracy of building extraction. The network's performance was rigorously validated on two benchmark datasets: the Google Image dataset and the ISPRS Vaihingen dataset.

The results demonstrated a notable improvement in extraction accuracy, showcasing the effectiveness of combining CNN and transformer methodologies. Our study employs transformers for feature extraction in a similar spirit, focusing on leveraging their advanced capabilities. However, while DSFENet is designed for segmentation tasks, our objective is centered on image classification. By utilizing transformers, we

aim to achieve accurate and efficient classification of images, building on the principles demonstrated by DSFENet but applying them to a different aspect of image analysis. The approach underscores the versatility and potential of transformers in various applications, extending beyond segmentation to classification tasks.

Another recent and noteworthy study focused on incorporating a vision transformer for feature extraction was conducted for brain tumor classification (Abbasi et al., 2022), which plays a crucial role in prognosis and treatment planning. This research aimed to advance multiclass brain tumor classification by leveraging the vision transformer as a sophisticated feature extractor. Initially, the study involved optimizing a deep learning model, specifically ResNet101, for the task of feature extraction. Following this, various machine learning classifiers were employed to perform multiclass classification based on the extracted features. The research then progressed by optimizing the vision transformer model and integrating it into the classification pipeline as an input to a machine-learning decision classifier. This approach allowed for a refined and enhanced classification process, demonstrating the potential of vision transformers in improving the accuracy and effectiveness of classification tasks. Our work aligns closely with this study, as we also focus on utilizing vision transformers for feature extraction.

However, instead of brain tumors, our objective is to apply these advanced feature extraction techniques to the multiclass classification of colorectal polyps. By leveraging the capabilities of vision transformers, we aim to achieve precise and reliable classification results, similar to the advancements made in brain tumor classification, but tailored to the specific challenges of colorectal polyp analysis.

A significant contribution utilizing a vision transformer for feature extraction

was presented by Gu et al. (2023), who developed an advanced model for classifying Diabetic Retinopathy (DR) from fundus images. This innovative approach effectively identified and categorized all stages of Diabetic Retinopathy, including no DR, mild, moderate, severe, and proliferative forms. The model is structured into two key modules: the Feature Extraction Block (FEB) and the Gradual Prediction Block (GPB). In the FEB module, a vision transformer is employed to extract features from fundus images. This module incorporates advanced attention mechanisms, allowing the model to focus more precisely on critical areas such as retinal hemorrhages and exudates. By leveraging these finer attention capabilities, the FEB module enhances the model's ability to identify and distinguish between the different stages of DR. The GPB module, on the other hand, is responsible for classifying the five stages of DR. It uses residual attention mechanisms to refine the classification process further, combining with the FEB module to capture a broad range of spatial features associated with each class of DR.

The study highlights the effectiveness of using vision transformers for detailed and nuanced feature extraction in multi-class classification tasks. Given these advancements, incorporating a similar approach in our work on multiclass colorectal polyp classification could offer substantial benefits. By utilizing the vision transformer's ability to capture intricate and diverse features, our model could achieve enhanced accuracy and reliability in distinguishing between different types of colorectal polyps.

Jiseob et al. (2023) introduced an innovative approach to Generalized Zero-Shot Learning (GZSL) that significantly enhances attribute-related information within image features by leveraging a ViT. This approach capitalizes on the Vision Transformer's ability to process the entire image region without degrading the image resolution, thereby

maintaining the integrity of local image information within the patch features. In traditional image processing models, preserving high-resolution details can be challenging, but the ViT's architecture addresses this by treating image patches as discrete tokens, which are processed with attention mechanisms.

The study harnesses both patch and Classification Token (CLS) features from the Vision Transformer to extract detailed and comprehensive image attributes. This method enables the model to capture and utilise nuanced visual information that might be lost in other models, leading to more accurate and effective learning of attributes. By leveraging these features, the authors were able to improve the performance of Generalized Zero-Shot Learning, a framework that traditionally struggles with accurately classifying unseen classes due to limited attribute-related information. The approach underscores the benefits of utilizing Vision Transformers for complex tasks such as GZSL, where detailed attribute extraction is crucial. By integrating the patch and CLS features, the model effectively enhances its understanding of image attributes, thereby advancing the field of zero-shot learning and demonstrating the Vision Transformer's potential in addressing challenges associated with learning and classifying new and unseen image categories. This innovative use of ViT in GZSL offers valuable insights for improving performance in various image classification tasks.

Conv-ViT, a novel hybrid model, was developed to enhance the detection of retinal diseases through the analysis of foveal-cut optical coherence tomography (OCT) images (Dutta et al., 2023). This innovative model integrates the strengths of both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to achieve superior performance in retinal disease classification. Specifically, the Conv-ViT model employs the CNN architectures Inception-V3 and ResNet-50 for detailed pattern analysis

through transfer learning, which focuses on understanding pixel correlations within the images. These CNNs excel in capturing local features and spatial hierarchies, making them well-suited for analyzing intricate patterns and textures in OCT images.

In addition to the CNN components, Conv-ViT incorporates a Vision Transformer model to leverage its ability to capture shape-based characteristics and analyze correlations among distant pixels. Unlike traditional CNNs that primarily focus on local pixel relationships, the Vision Transformer is designed to process global context by examining the entire image and understanding long-range dependencies. This allows the Conv-ViT model to effectively learn and integrate both local and global features, enhancing its capability to distinguish between different retinal conditions. By combining the CNNs' prowess in pixel-level pattern analysis with the Vision Transformer's strength in global feature extraction, the Conv-ViT model achieves improved classification accuracy across four retinal disease categories: choroidal neovascularization (CNV), diabetic macular edema (DME), DRUSEN, and normal. The study not only demonstrates the effectiveness of this hybrid approach in retinal disease detection but also highlights the efficiency and robustness of the Vision Transformer in feature extraction tasks. The successful integration of these diverse models underscores the potential of hybrid architectures in advancing medical image analysis and disease classification.

2.4 Model Explainability

An advanced automated method was developed for identifying and classifying various subtypes of lung and colon cancers using histopathology images from the LC25000 dataset (Tummala et al., 2023). This method encompasses the classification of different

lung cancer subtypes, such as adenocarcinoma, benign lesions, and carcinoma, as well as colon cancer subtypes, including benign tumors and adenocarcinoma. To achieve this, the authors proposed a sophisticated approach incorporating Grad-CAM (Gradient-weighted Class Activation Mapping) to generate detailed visual saliency maps. These maps are instrumental in accurately identifying and highlighting the critical regions within tissue pathology images that the model focuses on when making predictions about cancer subtypes. Using saliency maps increases the understandability of the presented models by indicating on tissue images the areas which were important for the model-making decisions. This allows pathologists to understand the exact areas which the model is emphasizing, an important step in the process of confirming and improving diagnostic procedures. Their role can also enhance the understanding of how the automated diagnosis works and the regions of interest for the model, thus assisting the pathologists in constructing more efficient and individualized treatment options.

The proposed pipeline aims to fully automate the detection and classification of lung and colon cancers from tissue pathology samples, integrating an explanation component to make the results more interpretable and actionable. This makes the method particularly well-suited for deployment in clinical settings, where the ability to automate complex diagnostic tasks while providing clear, understandable explanations can significantly enhance the efficiency and accuracy of cancer detection and treatment planning.

A sophisticated technique for improving the explainability of deep neural network (DNN) models has been developed through the use of synthetic histology images generated by a cGAN (Dolezal et al., 2023). This innovative approach leverages the

capabilities of cGANs to produce high-quality histology images that significantly enhance the performance of DNN architectures trained for classifying molecularly subtyped tumors. By utilizing synthetic histology, the technique reveals crucial histologic features associated with different molecular states of tumors, which might otherwise be difficult to discern. The procedure pays attention to perfecting enhanced artificial histology images using methods like class and layer blending. This emphasis allows for greater scrutiny of and differentiation between minor morphological changes within the different tumor subtypes. The method, therefore, enhances the quality of these synthetic images, thus making it possible to recognize crucial nuances in tumor types that make classification and diagnosis accurate.

Furthermore, the researchers illustrate how synthetic histopathology can be effectively used to enhance the education of pathologists-in-training. The intuitive and detailed visualizations provided by synthetic histology offer valuable insights into the histological characteristics of tumor biology. These visualizations can serve as powerful educational tools, helping trainees to better grasp the complexities of tumor morphology and the molecular features associated with different cancer types. This approach not only aids in the training of future pathologists but also contributes to a deeper and more comprehensive understanding of histopathological features in cancer diagnosis and treatment.

Mukhtorov et al. (2023) introduced an innovative and interpretable deep learning (DL) approach that integrates ResNet152 with Grad-CAM specifically for analyzing endoscopic images. Their method leverages the open-source KVASIR dataset, which includes a substantial collection of 8,000 wireless capsule endoscopy images. This dataset provides a rich source of data for training and validating the model, enhancing its ability

to classify and interpret various endoscopic findings. The integration of Grad-CAM (Gradient-weighted Class Activation Mapping) with ResNet152 allows for the generation of heatmaps that visually represent the areas of the endoscopic images that significantly influence the model's classification decisions. This approach not only improves the transparency of the model's decision-making process but also provides valuable insights into which regions of the images are most relevant for identifying specific conditions or abnormalities.

In addition to utilizing Grad-CAM for interpretability, the researchers applied a highly effective data augmentation technique. Data augmentation helps expand the variety and volume of the training data, contributing to the robustness and generalisation of the model. The combination of these techniques resulted in remarkable performance outcomes. The model achieved an impressive training accuracy of 98.3% and a validation accuracy of 93.5% in biomedical image classification tasks. These high accuracy rates highlight the effectiveness of the proposed approach in accurately classifying endoscopic images and demonstrate its potential for improving diagnostic accuracy in clinical settings. In this paper, the author elucidated the mathematical framework underlying the proposed classifier, clarifying that it is designed not to serve as a fully automated diagnostic tool but rather as a supportive system for medical professionals (Sabol et al., 2020). This distinction is crucial as it highlights the classifier's role in augmenting the expertise of pathologists rather than replacing their judgment. The study provides detailed insights into the mathematical algorithms and computational processes that drive the classifier, offering a comprehensive understanding of how it operates and contributes to the diagnostic workflow.

The research also includes a thorough evaluation of the classifier's performance

using real-world CRC histopathological data. This evaluation is pivotal as it tests the system's accuracy and reliability in practical scenarios, ensuring that it can perform effectively under real clinical conditions. The classifier was put through rigorous clinical trials involving fourteen experienced pathologists, who assessed its suitability and utility in a real-world setting. These trials were instrumental in verifying that the system not only meets but also aligns with the high standards required in medical diagnostics. The findings demonstrate that the proposed method achieves accuracy levels comparable to those of state-of-the-art neural network models. This indicates that the classifier is not only technically proficient but also practically valuable in clinical practice. Moreover, the system has been favorably received by medical experts, who have recognized its potential as a diagnostic tool. This acceptance by pathologists underscores the classifier's practical utility and its role in enhancing the diagnostic process, contributing to more accurate and efficient cancer detection and diagnosis.

More recent publications describe more advanced Grad-CAMs, capturing the detail and precision at higher levels than others, but we chose to employ Grad-CAM in our work because of its proven sturdiness and reliability in numerous research efforts. There are clear means of improving neural networks whose decision processes are interpreted through visual means and Grad-CAM Qualias focuses on this very task. It appeals to works where issues concerning the understanding and transparency of models are key and is one of the goals of our research.

Moreover, the newer variants of Grad-CAM, such as Grad-CAM++, introduce additional computational complexity and require longer processing times. While these advanced versions offer improvements in precision and detail, the increased complexity and resource demands associated with them pose significant challenges. For instance,

Grad-CAM++ incorporates additional mechanisms to refine the saliency maps, but these enhancements come at the cost of higher computational overhead. This makes them less suitable for our study, where the focus is on balancing accuracy with practical considerations such as computational efficiency and processing time. In summary, while advanced Grad-CAM variants provide more refined visualizations, the decision to use standard Grad-CAM in our study is based on its established effectiveness and efficiency. The simplicity and reliability of Grad-CAM align well with the goals of our research, ensuring that we can achieve meaningful insights into model behaviour without incurring excessive computational costs.

2.5 Feature Selection

A robust unsupervised feature selection method was introduced, which integrates latent representation learning to enhance the process of identifying relevant features (Tang et al., 2019). This innovative approach shifts the focus from the original data space to the learned latent representation space for feature selection. By operating in this transformed space, the method achieves greater resilience to noise and variability, as it evaluates feature importance based on the more stable latent representations rather than the potentially noisy original data. The core of this method involves creating an affinity matrix that captures the relationships between different data instances. This matrix serves as a foundation for understanding how data points relate to each other in the latent space. To model these latent representations effectively, non-negative matrix factorization (NMF) is employed. NMF decomposes the data into non-negative factors, which aids in uncovering the underlying structure of the data and facilitates the learning of meaningful features.

Additionally, the method incorporates a regularisation term based on the graph to maintain the manifold structure of the original data space within the transformed feature space. This graph-based manifold regularisation ensures that the local geometric properties and relationships of the original data are preserved even after transformation. By doing so, the approach maintains the integrity of the intrinsic structure of the data while performing feature selection, leading to more accurate and reliable results. In summary, this feature selection method leverages latent representation learning and graph-based manifold regularisation to enhance robustness and accuracy. By focusing on learnt representations and preserving the manifold structure, key features are effectively identified in a manner that is resistant to noise and variability, ultimately improving the quality and interpretability of the feature selection process.

A comprehensive study on multi-view feature selection, particularly for social media analysis-based applications, was conducted (Yangxi et al., 2016). In this research, the authors proposed a novel method known as Multi-View Regularized Multi-View Feature Selection (MRMVFS). This approach integrates four distinct types of information to enhance the feature selection process: label information, label relationships, data distribution, and correlations among different views. By synthesizing these diverse sources of information, MRMVFS effectively identifies the most representative features from the original multi-view datasets, thus improving the quality and relevance of the selected features. The researchers introduced manifold regularisation into the multi-task feature selection framework, leveraging both the limited labeled samples and the larger pool of unlabeled samples. This technique involves using a regression model to exploit the label information available from the labeled samples, enhancing the feature selection process.

The inclusion of manifold regularisation helps in preserving the intrinsic structure of the data while utilizing both labeled and unlabeled samples to refine the feature selection. MRMVFS was evaluated on several benchmark datasets, including NUS-WIDE, Flickr, and Animal image datasets. The results demonstrated that MRMVFS outperforms other existing feature selection methods, showcasing its effectiveness in selecting relevant features for image annotation tasks. Although the primary focus of this study was on image annotation, which differs from our goal of image classification, the principles underlying MRMVFS are highly relevant to our research. In particular, the use of manifold learning for feature selection and model regularisation aligns with our objective of simultaneously performing feature selection and enhancing model performance through advanced learning techniques.

A novel approach known as Multi-Label Feature Selection via Manifold regularisation and Dependence Maximization (MRDM) was introduced for advanced feature selection tasks (Rui & Wu, 2021). The MRDM method begins by projecting the original data onto a low-dimensional manifold space that retains the same structural characteristics as the original data while incorporating dependencies related to class labels. This projection is crucial for preserving the inherent relationships within the data and enhancing the relevance of the features selected. In the MRDM framework, the feature selection process is carried out using an iterative algorithm. This algorithm leverages $l_{2,1}$ norm regularisation to achieve sparse coefficients, which are essential for effective multi-label feature selection. By applying this regularisation technique, MRDM ensures that only the most relevant features are retained, thus improving the model's performance in handling multi-label classification problems.

Several comprehensive experiments were conducted on various multi-label

datasets to assess the performance of the MRDM method. The results demonstrated that MRDM outperforms several state-of-the-art multi-label feature selection methods, showcasing its effectiveness in selecting pertinent features while maintaining a strong dependence on class labels. This approach not only enhances feature selection but also contributes to more accurate and efficient multi-label classification, making it a valuable tool in the field of machine learning and data analysis.

A novel feature selection framework, which transforms shallow linear embeddings into deep embeddings, was proposed by Dornaika (2020). The authors argue that conventional shallow embedding methods fall short in terms of enhancing the performance of learning tasks, particularly when it comes to more complex data representations. To address this limitation, the proposed method employs a recursive embedding strategy coupled with a selection scheme to significantly improve performance. This approach is inspired by advances in deep learning techniques and involves the construction of a cascade of embeddings. In this framework, the cascade consists of stacking both linear and nonlinear functions to generate a robust representation or decision. The process involves ranking the features of the data representation and eliminating those deemed less relevant. By creating a cascade-based structure, the approach effectively transforms noisy descriptors into relevant and condensed features through a linear manifold learning process. This transformation is achieved via feature subset selection, which refines the data representation to enhance the overall quality and effectiveness of the feature selection process.

This research bears similarity to our work in that both employ recursive embedding within a hierarchical feature extraction structure. However, while the proposed method utilises linear manifold learning techniques, our study distinguishes itself by

incorporating nonlinear manifold learning techniques prior to the classification phase. This distinction is crucial as nonlinear manifold learning can capture more complex relationships within the data, potentially leading to more accurate and effective classification outcomes.

A hybrid method for multi-label feature selection (LMFS) was introduced by Y. Zhang et al. (2022), which integrates logistic regression, manifold learning, and sparse regularisation techniques. The authors highlight a common shortcoming in existing models: they frequently overlook either the sample matrix or the label matrix. In their view, considering both matrices is essential for effective feature selection. Their approach addresses this gap by combining both the feature manifold and the label manifold in the construction of the multi-label feature selection model. The LMFS model is developed by incorporating logistic regression to merge these manifolds, enabling the model to leverage the strengths of both sample and label information. This integration allows the model to capture the complex relationships and dependencies between features and labels, thereby improving the accuracy of feature selection in multi-label settings. Additionally, the model employs an L2,1-norm sparse constraint, which contributes to building a more robust and efficient feature selection framework by enforcing sparsity in the solution. This constraint helps in identifying the most relevant features while disregarding those that are less important, thus refining the overall performance of the feature selection process. LMFS model integrates these methodologies. The limitation of the previous models regarding multi-label feature selection is overcome and improvement in the ability to select meaningful features in difficult datasets is realized. This hybridization is special in that it incorporates both manifold learning and sparse Regularisation in this case making it an important development in features selection.

A feature selection method known as Iso-GA was developed to enhance the accuracy of cancer classification (Wang et al., 2023). This method specifically addresses the challenges posed by the inherent randomness associated with feature selection algorithms. Iso-GA is designed to select a reduced subset of critical genes by taking into account the nonlinear geometry present in microarray gene expression data. To achieve this, Iso-GA integrates a nonlinear dimensionality reduction algorithm, Isomap, with a genetic algorithm (GA). This hybrid approach is intended to capture the complex nonlinear structure of gene expression data effectively. The Iso-GA method combines the dimensionality reduction capabilities of Isomap with the optimization strengths of GA, creating a framework that not only reduces the number of features but also maintains or improves classification accuracy. By focusing on the nonlinear aspects of the data, Iso-GA can select a smaller set of genes while achieving similar or better results compared to traditional methods. This capability helps in mitigating the randomness and variability that often affect metaheuristic algorithms used in feature selection.The results obtained using the Iso-GA framework indicate that it can effectively reduce the number of genes needed for classification without compromising accuracy. This demonstrates the method's potential in overcoming the randomness associated with feature selection processes and highlights the effectiveness of Isomap in producing clearer outcomes through dimensionality reduction. While the Iso-GA method is aligned with the goals of our study in terms of improving classification accuracy, it focuses on microarray gene expression data, whereas our research is centered around image analysis. This distinction underscores the applicability of different feature selection techniques depending on the type of data being analyzed.

2.6 Model regularisation and Image Classification

A novel method for mammogram classification that integrates Random Forest and Locally Linear Embedding (LLE) algorithms for extracting texture features was introduced (Xuejun et al., 2023). This method is structured in three distinct stages. The first stage involves preprocessing, aimed at enhancing contrast and reducing noise within the Region of Interest (ROI) images. In the second stage, texture features are extracted from the Grey Level Co-occurrence Matrix, resulting in a sixteen-dimensional feature set that serves as the input dataset for the LLE algorithm. The final stage involves mapping this data into a five-dimensional subspace to facilitate classification. In our work, we also have LLE to perform dimensionality reduction, paralleling the approach taken in this study. However, instead of employing Random Forest for classification, as done in the proposed method, we have opted for CNNs. This choice reflects a shift towards leveraging advanced neural network architectures to potentially enhance classification performance beyond traditional machine learning models.

Currently, the majority of image classification tasks are achieved through supervised learning methods. However, this approach faces significant challenges due to the difficulties in annotating high-quality datasets, and real-world datasets often exhibit complex nonlinear structures. The cost of manual annotation grows exponentially with the number of targets and the complexity of the images, creating substantial barriers to effective and scalable image classification. In light of these challenges, there is a growing interest in unsupervised image classification as a viable alternative. Traditional unsupervised classification techniques, which rely on Euclidean distance metrics, often struggle to capture the nonlinear structures inherent in many datasets. This limitation results in lower accuracy and less effective classification outcomes.

To address these issues, (Situ, 2023) proposed a novel two-step approach designed to improve unsupervised classification. The first step involves employing manifold learning techniques to extract and understand the nonlinear structure of the dataset. This step aims to uncover the intrinsic geometry of the data, which is crucial for accurate classification. In the second step, the authors use agglomerative clustering to generate pseudo-labels based on the extracted features. These pseudo-labels retain the mathematical structure of the original data with high precision. The process of obtaining these labels allows for the training of a neural network without the need for costly manual annotation. By leveraging these pseudo-labels, an unsupervised image classifier can be trained effectively, even on smaller datasets. This approach not only reduces the reliance on extensive manual labeling but also enables the application of advanced neural network models to large-scale datasets, improving classification accuracy and efficiency in real-world applications.

Tan & Jeong (2023) proposed a hybrid method that combines manifold learning with contrastive learning to enhance histopathology image classification. This innovative approach integrates manifold geodesic distance into the contrastive learning framework, aiming to improve feature representation for nonlinear feature manifolds. Unlike traditional contrastive learning methods that use Euclidean distances to represent similarities and differences among features, this novel method leverages geodesic distance, which more accurately captures the complex geometric relationships inherent in nonlinear data structures. The proposed method incorporates sub-classes and prototypes in a manner similar to conventional contrastive learning. However, it distinguishes itself by using geodesic distances to create more accurate sub-classes that better reflect the nonlinear

nature of the feature manifolds. This approach facilitates improved separation of features, resulting in larger margins between different classes and thus better classification performance.

To evaluate the effectiveness of this hybrid method, the authors conducted a series of experiments involving a control group and two manifold learning groups. The control group employed standard methods without manifold learning, while the manifold learning groups used LLE and Isomap algorithms to extract nonlinear structures from the data. Pseudo-labels were generated based on these extracted features, and neural networks were subsequently trained on these labels. The performance of the hybrid method was then compared against the control group and the manifold learning groups to assess improvements in classification accuracy. The results demonstrated that the integration of manifold learning with contrastive learning significantly enhanced feature representation and classification performance, validating the efficacy of the proposed approach in dealing with the complexities of histopathology images.

The authors found that the accuracy rates for the two groups that manifold learning algorithms to extract nonlinear structures were significantly higher compared to the control group, which did not use manifold learning. Specifically, the group that employed the Isomap algorithm achieved an impressive accuracy rate of 85% on the test set. This substantial improvement highlights the effectiveness of nonlinear manifold learning techniques in enhancing classification performance by better capturing the complex, high-dimensional relationships within the data. The results underscore the utility of manifold learning approaches, such as Isomap and LLE, in providing more accurate and nuanced feature representations, which are particularly beneficial for tasks involving intricate and nonlinear data structures. Given the demonstrated efficiency and

superior performance of these manifold learning methods, it is evident that they are well-suited for our research on the dimensionality reduction of colonoscopy images. By leveraging nonlinear manifold learning techniques, we can achieve more precise and reliable feature extraction, thereby improving the overall accuracy and effectiveness of our classification models for colonoscopy image analysis.

A novel approach known as Manifold-GCN was proposed for image classification in semi-supervised scenarios, particularly when dealing with limited labeled data (Pascotti Valem et al., 2023). This innovative method integrates Graph Convolutional Networks (GCN) with manifold learning techniques to enhance classification performance. In their approach, deep features are first extracted from images using transfer learning techniques with CNNs and ViT models. These deep features are then processed through unsupervised manifold learning algorithms, which rank structures based on these extracted features. The primary goal of manifold learning is to utilise the intrinsic manifold structure of the data for improved distance and similarity computations. In the Manifold-GCN method, recent unsupervised manifold learning techniques are employed to refine similarity measures through rank-based formulations. This helps in leveraging the inherent geometric structure of the data to enhance the performance of the classification model, even with limited labeled data. The research presented is somewhat analogous to our study in that it applies manifold learning to learned features. However, while their approach integrates a graph neural network, our study utilises CNNs for feature extraction and classification. Both approaches aim to leverage advanced learning techniques to improve classification accuracy in challenging scenarios, albeit through different methodologies.

A method designed to enhance the diagnosis process by automatically classifying the malignancy or benignity of lung nodules based on 3D CT image patches was developed (Ren et al., 2020). Recognizing that real-world data often exhibit an intrinsic structure embedded within a low-dimensional manifold, the authors aimed to address this by introducing an innovative approach known as Manifold Regularized Classification Deep Neural Network (MRC-DNN). This method leverages the manifold structure of the data to improve classification accuracy. The MRC-DNN approach involves directly classifying lung nodule images using their manifold representation, which simplifies the data structure into a more manageable form. By incorporating manifold regularisation, the network's training process is constrained to prevent overfitting. This regularisation technique ensures that the learned model remains robust and generalizes well, even with complex and high-dimensional data. The key advantage of this approach lies in its ability to maintain the intrinsic data structure while performing classification. The manifold representation helps in capturing the underlying patterns and relationships within the data, leading to more accurate predictions. Furthermore, the use of manifold regularisation helps in mitigating the risk of overfitting, thereby improving the reliability of the model in real-world diagnostic applications.

A study explored the hypothesis that CNNs could achieve enhanced feature learning by incorporating manifold learning into their regularisation process (Samad & Sekmen, 2021). Manifold learning algorithms are designed to identify a new feature space that preserves the geometric structure of high-dimensional data when reduced to lower dimensions. By transforming the data into this new embedding space, the representation of the data can become more distinctive and better suited for separation by linear classifiers or clustering based on Euclidean distance metrics. The authors of this

study investigate the potential of using nonlinear manifold learning as an auxiliary tool to complement deep model parameter learning. Their approach involves regularizing the CNN's objective function with manifold learning techniques, which helps in improving the quality of the learned features and enhances the overall performance of image classification tasks. This method aims to refine the feature space, making it more suitable for classification and clustering tasks by better preserving the underlying data structure. In line with this research, our work also incorporates model regularisation using manifold learning techniques prior to the final classification step. By leveraging manifold learning, we aim to improve feature representation and enhance classification performance, similar to the approach taken in this study. This alignment underscores the effectiveness of combining manifold learning with deep learning models to achieve more accurate and robust classification outcomes. A review of these of these techniques is given in Table 2.2.

2.7 Problem Formulation

This study underscores the importance of effective colorectal polyp image classification, specifically addressing challenges related to data insufficiency and class imbalance by generating highly concordant synthetic images. Our approach not only involves optimizing transformer models for feature extraction but also addresses the need for explainable models in the CRC domain. By capturing the intrinsic geometrical patterns of data and incorporating regularisation techniques, we aim to enhance model performance and robustness. Our strategy includes generating synthetic images to mitigate the issue of limited training data, thereby enabling more effective training and feature extraction. This synthetic data will facilitate the development of a model that not only achieves high

Table 2.2: Review of cGAN, FEx, XAI and Manifold Learning in medical diagnosis

Publication	Method	Dataset	Modality	Task	Evaluation Metric	Highlights
Synthetic Image Generation						
Sasmal et al. (2023)	GAN	UniToPatho	Histopathology	Data Augmentation	Accuracy: 0.87 (25%), 0.76 (50%)	Learn the distinguishing features in an adversarial manner from a limited amount of labelled data and a large amount of unlabeled data.
Raju & Rao (2022)	cycle GAN	Colorectal Histology dataset MNIST	Histopathology	Data Augmentation	Accuracy: 0.99 Precision:1.0 Recall:0.99 F1 score:0.99	Address data imbalance issues using a CycleGAN-based data augmentation technique.
Yoon et al. (2022)	GAN	Private Dataset	Colonoscopy	Data Augmentation	Sensitivity Specificity F1 score AUROC	Capture the morphological characteristics of SSLs to identify missed polyps.
Chen et al. (2021)	ECGAN	CIFAR-10 Tiny ImageNet ImageNet	Image Data	Data Augmentation	CIFAR-10: FID: 7.9, IS: 10.0 Intra-FID: 41.3 Tiny ImageNet: FID: 18.9, IS: 18.4 Intra-FID: 203.3	Show that classifiers can enhance the performance of cGANs.
Yu & Zhang (2020)	DCGAN LSGAN CapGAN	BrainWeb	MRI	Data Augmentation	Kullback Leibler: 1.16 Divergence: 1.13 Accuracy: AUC:	Develop an image representation that remains consistent despite changes in object pose and spatial relationships.
Rau et al. (2019)	Extended pix2pix	Phantom data	Colonoscopy	Data Augmentation	Mean L1-error: 1.41 Mean relative L1-error: 24.7 Mean RMSE: 1.65	Transform monocular endoscopic images to depth.
Shin et al. (2018)	cGAN	CVC-CLINIC	Colonoscopy	Data Augmentation	Precision: 79, 85.3 Recall: 75, 68.1	Generate realistic polyp images while maintaining the original image structures.
Feature Extraction and Model Explainability						
Jiseob et al. (2023)	ViT	AWA2 CUB-200-2011 SUN	Images	Feature Extraction	Accuracy: AWA2: 0.84, 0.63, 0.72 CUB: 0.78, 0.59, 0.67 SUN: 0.46, 0.51, 0.48	Enhance the attribute-related information within the image features.

Publication	Method	Dataset	Modality	Task	Evaluation Metric	Highlights
Dutta et al. (2023)	ViT	Mendeley	OCTID	Feature Extraction	Accuracy:0.94,0.92 F1 Score:0.94,0.92	Analyze shape-based features by assessing the correlation between distant pixels.
Gu et al. (2023)	ViT	IDRiD	Fundus images	Feature Extraction	Sensitivity:0.81 Specificity:0.82 Accuracy: 0.82 AUC: 0.90	Focus on retinal haemorrhage and exudate areas by extracting fine-grain details.
Xia et al. (2023)	ViT	Google Beijing ISPRS Vaihingen	Remote sensing image datasets	Feature Extraction	F1 Score:0.87	Dual-stream feature extraction network for precise building extraction.
Tandon & Namboodiri (2022)	Transfo -rmer	FVC 2000, 2002, 2004, 2006	Fingerprint images	Feature Extraction	FRR%@FAR = 0.1% ROC curves	Develop a time and memory-efficient method to extract both global and local representations of a fingerprint.
Abbasi et al. (2022)	ViT	BrainTumor Retrieval	MRI	Feature Extraction	Accuracy: 0.99 AUC: 1.00	Implement multiclass brain tumour classification using a Vision Transformer as a feature extractor.
Tummala et al. (2023)	Grad-CAM Visual Saliency Maps	LC25000	Histopathology	Explainability	Accuracy: 0.99 AUC: 0.9 F1 score: 0.99 BA: 0.99 MCC: 0.99	Accurately identify critical regions in histopathology images.
Dolezal et al. (2023)	cGAN	TCGA CPTAC CPTAC-LUAD CPTAC-LSCC	Histopathology	Explainability	Accuracy: 0.79, 0.83, 0.75	Enhance the interpretability of DNN models using cGANs for educating pathologists-in-training.
Mukhtorov et al. (2023)	Grad-CAM	KVASIR	Endoscopy	Explainability	Accuracy: 0.98	Enhance the interpretability of deep learning methods.
Sabol et al. (2020)	CFCMCC	HE stained CRC tissue slides	Histopathology	Explainability	Accuracy: 0.91 Precision: 0.91 Recall: 0.91 F1 score: 0.91	Develop an explainable classifier to aid inmedical decision-making.
Wang et al. (2023)	Isomap	Breast, colon, lung cancer	Microarray	Feature Selection	Micro-AUC: 0.8, 0.9, 0.9	Select the optimal subset of genes in microarray data

Publication	Method	Dataset	Modality	Task	Evaluation Metric	Highlights
Lv et al. (2021)	Manifold Learning	Mulan Library	Images	Feature Selection Model regularisation	mean: - 0.87, 0.78 - 0.82, 0.59 - 0.51, 0.73 - 0.47, 0.24 - 0.64	Integrate adaptive global structure learning and manifold learning (SFAM) to maintain both global and sparse reconstruction structures.
Rui & Wu (2021)	Manifold Learning	Mulan Library	Images	Feature Selection Model regularisation	Hamming loss: 8.4 Ranking loss: 5.8 Convergence: 5.4 One-error: 4.1, Avg Precision: 7.	Multi-label feature selection using manifold regularisation and dependence maximization (MRDM).
Dornaika (2020)	LDE	Yale Extended Yale PIE PF01	Image	Feature Selection	Accuracy: 0.95, 0.83, 0.91, 0.83	Transform a shallow embedding into a deep variant by cascading multiple layers of embeddings, where each layer undergoes feature selection and elimination.
J. Zhang et al. (2019)	Manifold Learning	Mulan Library	Images	Feature selection Model regularisation	Hamming loss: 32.58 Ranking loss: 14.49 Coverage 11.86 Avg precision: 19.79 Macro-F1: 21.39 Micro-F1 22.36	Incorporate manifold regularisation into multi-class label feature selection to enhance discriminative feature extraction.
Tang et al. (2019)	kNN	ORL, orlraws10P warpPIE10P, COIL20 Isolet, CLL_SUB_1116 Prostate_GE, USPS	Microarray	Feature Selection Model regularisation	ACA ± std: 0.78 ± 2.5, 0.82 ± 6.1, 0.91 ± 5.2, 0.96 ± 2.6, 0.90 ± 0.9, 0.60 ± 7.4, 0.91.07 ± 6.6, 0.93 ± 0.2	Integrate latent representation learning into unsupervised feature selection methods for enhanced feature extraction.
Yangxi et al. (2016)	kNN, SVM	NUS-WIDE Flickr Animal	Images	Feature Selection	N/A	utilise manifold regularisation in multi-view feature selection (MRMVFS) to leverage label information, label relationships, data distribution, and correlations among different types of features simultaneously, enhancing feature selection performance.

Publication	Method	Dataset	Modality	Task	Evaluation Metric	Highlights
Tan & Jeong (2023)	Pre-trained VGG16 with ImageNet	IHCCs Liver cancer	Histopathology	Classification	Accuracy: 0.77 Precision: 0.77 Recall: 0.76 F1 score: 0.76	Perform dimensionality reduction using geodesic-distance-based feature clustering to evaluate contrastive loss, instead of cosine-distance effectively.
Pascotti Valem et al. (2023)	GCN	Flowers17 Corel5k CUB-200-2011	Images	Classification	Accuracy: 0.97, 0.95, 0.79	Integrate Graph Convolutional Network (GCN) with manifold-based approaches for image classification.
Xuejun et al. (2023)	LLE	N/A	Mammogram	Classification	Avg Accuracy: 0.92 AUC: 0.99	A mammogram classification method that combines Random Forest with the Locally Linear Embedding (LLE) dimensionality reduction algorithm for texture features.
Situ (2023)	LLE	MNSIT	Images	Classification	Accuracy: 0.97	Extract the nonlinear structure of the original dataset using manifold learning methods, followed by generating pseudo-labels using the agglomerative clustering algorithm.
Samad & Sekmen (2021)	LLE	CIFAR Macular Disease Image	OCT	Classification	AUC: CIFAR:0.92 OCT:0.99	Examine the impact of regularizing CNNs with nonlinear embeddings of convolutional response features.
Ren et al. (2020)	Manifold	3D LIDC	CT	Model regularisation, Classification	Accuracy: 0.90 Sensitivity: 0.81 Specificity: 0.95	Enhance the classification performance of DNNs using manifold-based constraints to regularize the training process.

classification accuracy but also provides transparent and interpretable results. The goal is to build a model that pathologists can trust, given the complexity of the data involved. By concentrating on these critical aspects, our work highlights the robustness of the proposed framework in achieving both explainability and enhanced generalisability in the classifier. The integration of synthetic image generation, optimization of feature extraction processes, and application of model regularisation collectively ensure that our approach adeptly addresses the inherent challenges associated with data insufficiency and class imbalance.

Synthetic images are significant in enhancing the training dataset, as this mitigates the risk of bettering real images than what is available in reality. This enhances the scope of training as the model is exposed to larger number of different examples which translates to better feature learning. In addition, response optimization is done considering cutting edge methods particularly transformers, this leads to better performance since the model extracts the best features. In addition, various model regularisation schemes are also adopted to control overfitting and maintain models that are usable in different environments and datasets. This model is not only beneficial in enhancing the performance of the model but also the structure. In general terms, our approach embraces the establishment of a classification framework for the detection of colorectal polyps that is not only efficient and effective, but more importantly, clear and trustworthy. We seek to develop a workable solution that addresses the problem in modern-day medicine where the diagnosis and subsequent management of colorectal cancer are bedeviled by too many uncertainties.

Feature extraction is a pivotal step in the process of image diagnostics, playing a significant role in accurately identifying and classifying images. For colorectal polyp

feature extraction, several methods are currently employed, including end-to-end learning, hand-crafted features, and hybrid approaches (Bernal et al., 2017). These methods utilise various strategies to capture the essential characteristics of polyps in diagnostic images. End-to-end learning approaches involve training models that learn to extract features directly from the raw image data. This method leverages deep learning techniques to automatically learn and represent the complex patterns in the images, which can enhance the model's ability to identify subtle details relevant to polyp classification. Hand-crafted feature extraction methods rely on predefined algorithms to extract low-level features such as texture, structure, color, and magnitude from the images. These features are manually designed and chosen based on domain expertise and prior knowledge. While these methods can provide valuable insights, they may not always capture the intricate details required for accurate polyp classification.

Hybrid methods combine elements of both end-to-end learning and hand-crafted features. By integrating the strengths of automated learning with manually designed features, these methods aim to provide a more comprehensive representation of the polyp characteristics. Despite the advancements in these approaches, the challenge remains to extract complete and accurate features that include fine-grain details. Achieving this is essential for minimizing the misclassification rate and improving the diagnostic accuracy of colorectal polyp detection. Effective feature extraction techniques are crucial for developing reliable diagnostic tools that can significantly impact patient outcomes.

Both data availability and feature extraction occur in relation to the supply of Computer-Aided Detection (CAD) systems. In addition, these issues are complex and to an extent amenable to resolution. In the past, studies twinned either one of these problems and tried to resolve them separately, but it is much harder to embrace both

problems at once. The problem of limited data is quite common and a lot of authors cite an obvious solution to the problem of data: the use of data augmentation techniques. Basic augmentation methods are: rotation, flipping, brightness modifying, scaling and other related techniques. On the one hand, it is a good practice to extend the number of images by creating its variations, on the other hand it does not provide any additional new samples.

Therefore, the obtained augmented data are still subject to the weaknesses of the preliminary data set and fail to represent the diversity that may exist in reality. Such limitation in data diversity may greatly affect feature extraction processes. In the absence of new unseen samples, feature extraction methods might not perform adequately on a wider span of variations. Eventually, this can cause a decrease in a classifier's sensitivity and performance as a whole. The shortage of novel and diverse samples makes feature extraction algorithms less efficient at acquiring all the relevant, broad and therefore more useful features that would assist in enhancing the rigor and precision of CAD systems. It is important that both issues of data scarcity and difficulties in feature extraction are tackled to come up with better and dependable diagnostic tools.

Several strategies focus on enhancing classifier performance by employing various data augmentation techniques. One prominent method in this area is the use of Generative Adversarial Networks (GANs). GANs are sophisticated deep learning models that consist of two neural networks: a generator and a discriminator. These networks work in opposition, with the generator aiming to create realistic images and the discriminator striving to distinguish between real and generated images (Elsheikh, 2008). This adversarial process helps GANs produce high-quality synthetic images that can augment existing datasets. Conditional Generative Adversarial Networks (cGANs)

extend the functionality of GANs by incorporating additional information to guide the image generation process. This allows cGANs to create images that belong to specific classes (Krizhevsky et al., 2012) (Sener & Savarese, 2017). By leveraging this conditional information, cGANs can enhance the diversity of the training data, making it more representative of various classes and improving the robustness of the classifier.

In the context of image classification, Vision Transformers (ViTs) have shown significant promise due to their ability to capture long-range dependencies and handle complex contextual information effectively. ViTs excel at processing large images and extracting features from fine-grain details. However, the effectiveness of a Vision Transformer model is heavily dependent on the availability of a substantial dataset. A significant challenge in this domain is the availability of public colorectal cancer (CRC) datasets and the associated class imbalance issues.

To address these challenges, we have integrated a Conditional Deep Convolutional GAN (cDCGAN) into our framework to generate synthetic images. GANs, by introducing new and unseen images into the dataset, play a crucial role in increasing data diversity. On the other hand, cDCGAN focuses on generating images for specific class labels using additional information, which helps tackle the class imbalance problem prevalent in CRC data (Krizhevsky et al., 2012) (Sener & Savarese, 2017). Nevertheless, it is necessary to analyze the characteristics and level of variation of the synthetic images created by the cDCGAN network. However, prior to the overfitting of the classifiers due to augmentation fusion, extremely concordant images need to be extracted. To solve this problem, a further classifier is added to the network which classifies synthetic images according to the maximal predicted probability for each of available classes. To investigate this quality and variability of these samples, two powerful parameters Fréchet

Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Park et al., 2022) are employed. These parameters play an important role in ensuring that the synthetic images used do not deteriorate the training aimed at the classifier and its performance.

The black-box nature of neural networks, which is one of their most advanced capabilities, complicates the understanding of how and why systems make specific decisions. This issue is especially serious in medicine where it is essential to garner and keep pathologists' confidence in the systems. To overcome these issues, we have constructed an Explainable Artificial Intelligence model. The quick and easy access of the patient, as well as the analyzing, decided quality of the diagnosed conditions description will be further enhanced using our proposed XAI model. To make it clearer for the reader, the authors of this work have made the images associated with the model more readable using Grad-CAM technology. Grad-CAM is an example of an attention map that indicates which parts of the input image hold the most relevance to the model output. These maps efficiently illustrate all aspects of the input which were critical in making the predictions or classification. In our model, it is argued that the incorporation of Grad-CAM heatmaps works in closing the gap between DNNs complexities and the practicality of the medical domain. The pathologists can see the areas highlighted under the model and appreciate how the model behaves and what decisions it makes. Such validation of the model's outputs addresses the concerns on the models reliability and accuracy. Our method intends in the end that the automated solutions provided will not only be efficient but will also correspond to the needs of the healthcare practitioners, thereby improving their willingness to use them in clinical decision making.

An increased number of extracted features typically results in high-dimensional data, which complicates the task of capturing meaningful patterns among the spread-out

data points. This complexity often leads to notorious overfitting problems, significantly affecting the model's performance. To address this, it's crucial to build a generalized and robust model. In the context of image classification, where images often contain nonlinear and complex structures, linear techniques may prove inadequate. Consequently, nonlinear dimensionality reduction techniques are better suited for handling such data.

The manifold algorithm is designed to capture these characteristics in high dimensions and conserve them in newly constructed low-dimensional embeddings. While techniques like Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) are commonly used for dimensionality reduction, they rely on linear transformations, which are less effective at learning nonlinear data manifolds. Nonlinear dimensionality reduction methods, such as Isomap and LLE, are more adept at uncovering the underlying structure of high-dimensional data. Isomap focuses on preserving the geodesic distances between data points, while LLE excels at maintaining the local relationships between neighboring points. In this study, these algorithms are within the proposed framework to identify the most effective method for discovering the underlying nonlinear structure of CRC data. The aim is to integrate these techniques into a deep CNN-based classification framework to analyze colon endoscopy images, thereby addressing the challenges associated with high-dimensional data and improving classification accuracy.

2.8 Summary of Chapter

In this chapter, we have pinpointed the key discussions in relation to the literature and contributions such as this thesis. The literature retrials have demonstrated that many of

the described models lie in attempting to develop approaches to an imaging modality for detecting and classifying colorectal cancer polymorphic images. At the same time, there have been additions concerning this area of investigation, and several problems still exist, and they arise from the nature of CRC polyp data itself, which is why a single solution is insufficient.

The first threat described in the works is the publicly available gastrointestinal polyp databases. Due to the limited accessibility of such datasets, efficient classification models cannot be developed or tested. Aside from the data being scanty, the class imbalance problem of certain polyp subtypes not being sufficiently represented is a common one. Such a problem may result in biased models whose efficiency in predicting infrequent polyp types will be poor, and this can have dreadful consequences on the patients. To deal with these problems, conventional data augmentation methods have also been applied in previous work. These include techniques like rotation, flipping and scaling that expand the available training data. However, such conventional methods will have a major drawback: they do not add any new and previously unseen samples to the dataset. The data strengthened by these methods is also often not sufficient to train the model on the images it has not seen in the sense that it does not contain diversity. This is where more advanced augmentation techniques are required. One promising approach to overcoming this limitation is the use of conditional Generative Adversarial Networks (cGANs). cGANs have the capability to generate synthetic samples that are new and distinct from the original data, thereby enhancing the diversity of the dataset. However, the literature also points out a critical concern with the use of cGANs: the quality and concordance of the synthetic images. If the generated images do not accurately represent the characteristics of real CRC polyp images, they can degrade the performance

of the classifier rather than improve it. This highlights the importance of thoroughly evaluating the quality and diversity of synthetic images produced by cGANs before they are incorporated into the training process.

While significant advancements have been made in the detection and classification of CRC polyps, several unresolved challenges persist, particularly regarding the availability of datasets and the quality of synthetic data. Traditional data augmentation techniques, although useful, are insufficient for addressing the need for diverse and representative training data. cGANs offer a promising solution, but their effectiveness depends on the careful assessment of the synthetic images they generate. Thus, future work in this area must focus not only on generating synthetic data but also on ensuring that this data meets the necessary standards of quality and diversity to build intelligent and reliable classifiers. This chapter has set the stage for the methodologies discussed later in this thesis, which aim to address these critical challenges and contribute to the advancement of CRC polyp classification.

Secondly, the literature indicates that current methodologies for extracting colorectal polyp features primarily include end-to-end learning, manually crafted features, and hybrid approaches (Bernal et al., 2017). However, a critical gap remains in extracting comprehensive and precise features that capture the fine-grained details essential for minimizing the misclassification rate of polyps. These fine-grained details are crucial because colorectal polyps often exhibit complex morphological structures that, if not accurately captured, can lead to significant errors in classification. The prediction accuracy of CAD systems has been consistently hindered by two major challenges: the limited availability of high-quality data and the difficulties associated with effective feature extraction. While data augmentation methods, such as traditional image transformations,

help address the scarcity of images, they do little to enhance the diversity of the dataset. This lack of diversity ultimately restricts the ability to extract a wide range of features, particularly those that are subtle yet significant in differentiating between various polyp subtypes. Consequently, these limitations can lead to suboptimal performance of the classification model, especially in clinical scenarios where precision is critical.

Given these challenges, there is a pressing need for a more sophisticated feature extraction mechanism that can go beyond basic attributes like shape, texture, and colour. Such a mechanism must be capable of capturing the intricate and nuanced details present in colorectal polyp images, which are often characterized by complex and variable morphological features. An advanced feature extraction approach would not only improve the classifier's ability to distinguish between different types of polyps but also enhance its overall sensitivity and specificity. To develop a highly sensitive classifier model, it is essential to implement a feature extraction process that meticulously captures these fine-grain details. This process should be integrated into a comprehensive framework that also addresses the limitations of data augmentation by incorporating methods that introduce genuinely novel and diverse samples into the dataset. By doing so, the model can be trained on a dataset that is both sufficiently large and diverse, thereby improving its generalisability and reducing the likelihood of misclassification. While existing methods for colorectal polyp feature extraction have laid the groundwork, they fall short in addressing the need for detailed and precise feature capture, which is vital for improving the accuracy of CAD systems. The integration of advanced feature extraction techniques, coupled with strategies to enhance dataset diversity, is crucial for the development of a robust and highly sensitive classifier model that can reliably distinguish between different polyp subtypes and support more accurate and effective colorectal cancer screening.

Thirdly, another significant challenge faced by deep learning models, as highlighted by the literature, is their often opaque decision-making process. While the existing studies demonstrate the application of Explainable AI (XAI) within the medical domain, they frequently fail to integrate XAI as a fundamental component of the computer-aided diagnosis (CAD) system. This omission limits the ability of practitioners to gain insights into the internal workings and decision-making processes of deep neural network (DNN) models. Incorporating XAI into the CAD system can profoundly enhance transparency, allowing practitioners to understand how the model arrives at its predictions. Developing an interpretable CRC classification framework for colonoscopy imaging is, therefore, of paramount importance. Such a framework would not only clarify the specific features and factors that influence the model's predictions but also assist pathologists in making more informed and accurate diagnostic decisions. This increased interpretability can lead to greater trust in the system, ultimately improving clinical outcomes and fostering a more reliable integration of deep learning technologies into medical practice.

Fourthly, CRC polyp images often exhibit high dimensionality due to their intricate textures and detailed structures. This high dimensionality can exacerbate the notorious overfitting problem, where the model struggles to identify meaningful patterns within sparse or complex data, resulting in a model that lacks generalisability. To address this issue and promote better generalisation, it is crucial to apply regularisation techniques to the cost function of nonlinear dimensionality reduction algorithms. For instance, the manifold algorithm can be regularized to penalize overly complex models, thus mitigating the risk of overfitting. Despite the potential benefits of this approach, there has been limited research focusing on regularizing deep learning model training through manifold learning. To the best of our knowledge, there is a notable gap in

the CRC polyp domain regarding the enhancement of classification performance using such methods. This gap presents an opportunity for further exploration and innovation, highlighting the need for dedicated research to integrate manifold learning techniques with deep learning models in the context of CRC polyp classification.

The concerns previously discussed drove us to design a comprehensive methodology for CRC polyp image classification capable of addressing the multifaceted issues inherent in such images. Our approach integrates various components to tackle each challenge simultaneously, from data insufficiency and feature extraction to model interpretability and overfitting. The proposed methodology will undergo rigorous evaluation using advanced metrics to ensure high sensitivity and accuracy in classification. By developing an explainable deep learning method, we aim to build pathologists' trust and improve the transparency of the model's decision-making process. This enhanced transparency is crucial for integrating the model into clinical workflows, where it can provide clear insights into the reasons behind each prediction. Moreover, the methodology's focus on automating CRC biopsy workflows addresses key factors affecting pathologists' efficiency, such as the need for accurate and reliable support tools. By streamlining the biopsy process and enhancing the capabilities of CAD systems, the proposed framework not only saves pathologists time and effort but also improves diagnostic accuracy. This comprehensive approach ensures that the CRC polyp classification system is both robust and generalized, effectively advancing the state of automated medical diagnostics and contributing to more efficient and reliable clinical practices.

Chapter 3

Explainable Deep Learning Framework for CRC Polyp Classification

In this chapter, we present an overview of the proposed methodology for an efficient and explainable CRC classification model (XAI) for concordant images that facilitate the classification of polyp types in endoscopy images. As mentioned earlier, the main objectives of our study are to generate synthetic unseen images with strong concordance, to extract features using an attention-based mechanism and its feature discriminative abilities, to provide transparency for a reliable clinical decision-making process through an explainable model, to uncover the nonlinear structures present in the data for efficient model training along with model regularisation, and finally, to perform classification with optimised deep learning model.

3.1 Methodological Framework

The proposed framework comprises six key modules: Ensemble Learning Block - **ELB**, Synthetic Image Generation Block - **SIGB**, Deep Feature Extraction Block - **DFEB**, Model Explainable Block - **MEB**, Manifold Regularization Block - **MRB**, and Deep Neural Network - **DNN** for final classification. All these parts of the framework are described in the subsections 3.1.1 - 3.1.6 below.

3.1.1 Ensemble Learning Block (ELB)

The classification of colorectal polyps presents a complex and multifaceted challenge, particularly when attempting to achieve automated classification through deep learning networks. This complexity arises from the intricate and highly variable patterns these polyps exhibit, making distinguishing between different polyps, such as benign and malignant, especially difficult. Recognising these challenges, we have structured the first phase of our methodology around developing an ensemble learning model. This model is carefully designed to integrate various methods by combining multiple base learners, thereby harnessing a broader and more diverse range of features extracted from input data.

The rationale behind this approach stems from the observation that while powerful, individual CNN architectures do not consistently deliver outstanding performance on their own. Each CNN model has unique strengths and weaknesses, and performance can vary significantly depending on the specific characteristics of the data set used. Therefore, by combining the strengths of multiple weak learners, each potentially capturing different aspects or features of the polyps, we can construct an ensemble model that is more robust

and capable of achieving superior overall performance. This ensemble-based strategy is particularly advantageous for analysing colonoscopy images, where subtle and complex variations in the texture, shape, and colour patterns of colorectal polyps require a more nuanced and comprehensive approach.

Our proposed ensemble model leverages CNNs to analyse and classify these images. The key steps in this approach include several critical processes, including data augmentation, which involves artificially increasing the size and variability of the training dataset to improve model generalisation; hyperparameter optimisation, which fine-tunes the model parameters to enhance its performance; and finally, the ensemble-based classification of colorectal polyps, where multiple CNNs are combined to improve the accuracy and robustness of the predictions. By employing this ensemble approach, we aim to address the limitations of single-model architectures and provide a more reliable and accurate classification framework for colorectal polyp detection and diagnosis. This not only improves the effectiveness of CAD systems but also has the potential to significantly impact clinical practice by enabling more accurate and early detection of colorectal cancer, ultimately contributing to better patient outcomes and advancing the integration of AI technologies in healthcare.

In addition to its numerous advantages, one of the primary requirements for the effective implementation of deep learning models in medical imaging is the availability of large, high-quality medical datasets for automated model training. However, generating a sufficiently large and meticulously annotated training dataset presents significant challenges. This difficulty arises from the inherently expensive and time-consuming nature of data acquisition and annotation in the medical field, which often involves domain-specific expertise to accurately label complex medical images (Jha et al., 2021;

Nogueira-Rodríguez et al., 2021). The scarcity of large datasets is further compounded by privacy concerns and the difficulty of accessing diverse patient populations, making it particularly challenging to gather a dataset that is both representative and comprehensive.

To mitigate these challenges, transfer learning is commonly employed in conjunction with data augmentation techniques to maximise the utility of available data and leverage the power of pre-trained architectures (Pacal et al., 2020). Transfer learning is a powerful technique that is particularly suited for situations with insufficient training data to train a model from scratch. This approach involves utilising deep neural networks trained extensively on large, diverse datasets, such as ImageNet, and then transferring the learnt weights and knowledge to new tasks requiring classification with smaller, domain-specific datasets. By doing so, the pre-trained models bring forward a wealth of feature representations that can be fine-tuned for colorectal polyp classification, thereby improving model performance even with limited new data.

In addition, there is a significant imbalance in the distribution of colorectal polyp image data among various patients, which poses another major challenge for effective image classification. This imbalance can lead to biased models that perform poorly on under-represented classes. For instance, images of adenocarcinoma, a type of cancerous polyp with a carcinoma structure, are exceedingly scarce in the dataset. This scarcity can prevent the model from learning the distinct features associated with this critical class, potentially compromising the model's ability to accurately identify adenocarcinoma cases. To address this imbalance, data augmentation techniques such as image flipping, rotation, scaling, and brightness adjustments were employed. These methods help artificially increase the number of training samples in the underrepresented classes, specifically the hyperplastic and adenocarcinoma categories, thereby ensuring

a more balanced representation across all three classes in the dataset. In the Colorectal Polyp Dataset Collection (CPDC), these augmentation techniques were systematically applied to enhance the number of training samples available for the less frequent classes, promoting a more balanced and fair model training process.

Furthermore, when employing deep learning algorithms to classify colorectal polyps, it is crucial to carefully consider and optimise hyperparameter configurations to achieve the best possible model performance. Hyperparameters, such as learning rates, batch sizes, number of layers, and types of activation functions, play a vital role in the training process and significantly impact the model's ability to learn from data effectively. Fine-tuning these parameters can lead to substantial improvements in model accuracy, generalisation ability, and robustness. By systematically exploring different hyperparameter settings through methods such as grid search or random search and incorporating domain-specific knowledge, it is possible to identify the optimal configuration that maximises the model's performance in real-world clinical settings. This careful attention to hyperparameter optimisation ensures that deep learning models used for colorectal polyp classification are not only accurate and reliable but also adaptable to varying dataset sizes and conditions, ultimately contributing to more effective and trustworthy AI-driven diagnostic tools in healthcare. Hence, this part of the framework addresses **RQ1a** for optimal hyperparameter settings of the architecture.

Six contemporary CNN architectures were used to accomplish the hyperparameter tuning. The purpose of this step is to find the most suitable hyperparameter settings among neural networks for the purpose of classifications. First, the datasets were segmented into a training and a test set. Then the training set was passed to the pre-trained network to find out the best possible optimiser for the dataset. The three

optimisers selected are ADAM - Adaptive Moment Estimation, SGDM - Stochastic Gradient Descent, and RMSprop - Root Mean Square Propagation. During the first phase of the experiment, six networks were tested with these three optimizers one by one with a learning rate of 0.001 and 50 epochs. In the next step, the networks were tested for the same learning rate but by increasing the number of epochs from 50 to 100. Later, the bending ratio was increased to 0.005, and the number of epochs was set to 50 and then 100 for the next phase of the experiment. After establishing the most efficient optimiser, further experiments were conducted to select the number of epochs that generated the best result. The best resulting hyperparameters were used for further experimentation.

The rationale for employing ensemble learning in the classification of colorectal polyps is rooted in enhancing the diagnostic system's generalisation ability. In complex medical imaging tasks, such as the identification and classification of colorectal polyps, a single learning model or learner may encounter difficulties in fully capturing the underlying data distribution, especially given the variability and complexity inherent in medical images. Individual models often possess unique strengths and weaknesses, and these limitations can result in suboptimal performance when used in isolation. However, by aggregating the decisions of multiple weak learners, ensemble learning significantly bolsters the learning capability of the classification system. This approach helps to mitigate the individual limitations of each model, resulting in a more robust and comprehensive understanding of the data.

Ensemble learning works by consolidating the decisions of diverse learners, effectively combining their individual outputs to produce a more accurate and reliable final prediction. This process not only improves the robustness of the system but also enhances its overall performance by ensuring that the model is less likely to overfit

to any particular subset of the training data. In the proposed method, three cutting-edge deep learning models—GoogLeNet, Xception, and ResNet-50—were employed as foundational benchmarks. These CNN architectures are renowned for their ability to handle complex image recognition tasks due to their advanced structural designs, such as residual learning. Each model differs in layer configurations, block designs, and specific learning techniques. For instance, GoogLeNet utilises an Inception module for multiscale feature extraction, Xception leverages depthwise separable convolutions for efficient computation, and ResNet-50 employs residual blocks to mitigate the vanishing gradient problem during training.

Following the implementation of these deep learning architectures, an averaging approach was initially used to combine the predictions of these base learners. This simple ensemble technique aggregates the outputs of each model by calculating the average of their predictions, thus generating a final decision that reflects a consensus among the learners. In addition to simple averaging, a more sophisticated weighted averaging method was also implemented. This approach assigned appropriate weights to each base classifier based on their performance metrics, such as accuracy or F1-score. The weighted average aims to give more importance to models that perform better in certain aspects of the classification task, thereby enhancing the overall classification accuracy. This weighted ensemble strategy, as depicted in Figure 3.1, leverages the unique strengths of each model to optimise performance in the challenging domain of colorectal polyp classification. The effectiveness of ensemble learning is largely driven by the diversity among its base learners, which can be achieved through several strategies. First, diversity can be introduced by employing a range of different learning algorithms or by varying the configurations of the same algorithm. For example, using different architectures or

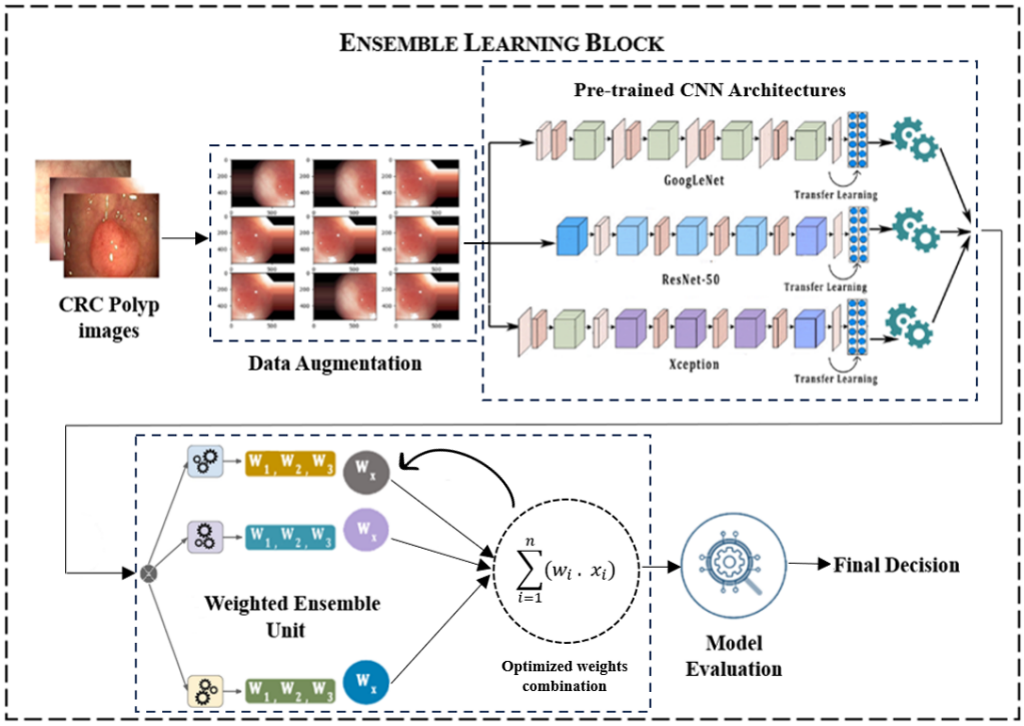

Figure 3.1: Methodology of weighted-average ensemble learning block

hyperparameter settings can lead to models that focus on different features or aspects of the data. Second, utilising a diverse set of features for training each base learner can help ensure that different models capture different relationships and patterns within the data. Finally, incorporating a variety of training instances or data subsets for each learner helps to expose the models to different aspects of the problem domain, thereby enhancing the ensemble's ability to generalise well across unseen data. These diverse approaches ensure that the ensemble captures a wide range of perspectives on the classification task, thereby improving its overall effectiveness and reliability (Gomes et al., 2017). By integrating these diverse models and techniques, the ensemble learning strategy in this study aims to create a more accurate and reliable system for the automated classification of colorectal polyps, ultimately advancing the field of CAD and contributing to improved

clinical outcomes.

Furthermore, the classification of serrated adenomas presents a particularly challenging task due to their hybrid nature, which differs significantly from the more distinct characteristics exhibited by hyperplastic and adenomatous polyps. Serrated adenomas have a combination of features that overlap with benign and potentially malignant polyps, making their identification and classification less straightforward. This complexity arises because the visual patterns of serrated adenomas can closely resemble those of other types of polyps, making them difficult to distinguish even for experienced pathologists. The subtle morphological differences between these polyps often require careful examination and interpretation, which can be both time-consuming and subject to variability among different observers. Therefore, it is crucial to develop a reliable and precise automated assessment system that can accurately classify serrated adenomas and other polyp types to ensure proper diagnosis and treatment planning.

The availability of numerous pre-trained architectures for transfer learning offers a potential solution to enhance the performance of automated classification systems. However, the efficacy of these pre-trained models must be carefully compared and evaluated to identify the architecture that consistently performs best for the specific task of colorectal polyp classification. Different pre-trained models, such as VGG, ResNet, Inception, and DenseNet, have unique strengths and weaknesses, particularly in handling the nuances of medical imaging data. The performance of these models can vary based on factors such as network depth, the complexity of the convolutional layers, and the types of features they are optimised to detect. As a result, a thorough evaluation and comparison of these architectures are essential to determine the most suitable model for the classification of colorectal polyps. As part of our comprehensive analysis, we

have conducted a detailed examination and assessment of various transfer learning architectures to address this need. This evaluation aims to determine which architecture is most appropriate for accurately classifying colorectal polyps, considering factors such as model accuracy, computational efficiency, and the ability to generalise from training to unseen data. By utilising pre-trained architectures instead of developing a customised deep neural network from scratch, we leverage the rich feature representations that these models have already learnt from vast datasets. This approach not only saves significant time and resources but also enhances the model's ability to generalise to new data, as these pre-trained models have already been fine-tuned to recognise a broad range of features that are applicable to various image classification tasks. The decision to use pre-trained architectures was informed by our need to address **RQ1b**, which focusses on optimising model performance for accurate and reliable colorectal polyp classification. The training scheme employed for this block of our methodology, which integrates these pre-trained architectures, is detailed in Algorithm 1. This algorithm outlines the specific steps and procedures used to fine-tune the pre-trained models, ensuring that they are optimised for the nuances and specificities of the colorectal polyp classification task, ultimately contributing to a more robust and effective computer-assisted diagnosis (CAD) system for clinical applications.

3.1.2 Synthetic Image Generation Block (SIGB)

In the second phase, our framework aims to produce high-quality synthetic images for model training to address **RQ1c**. To achieve this, it is beneficial to up-sample the data to a larger output feature map. To up-sample, we embed a transposed convolutional layer in the GAN instead of a regular convolutional layer, which typically down-samples the data.

Algorithm 1 The training scheme of ELB

Inputs:

1: Training data $Q = \{ q_1, q_2, \ldots, q_n \}$

2: Hyperparameter search space $P = \{ p_1, p_2, \ldots, p_n \}$

3: Number of trials *(X)*

4: Number of classes *(C)*

5: Number of models to be used in the ensemble *(N)*

6: Testing data $Q' = \{ q_1', q_2', \ldots, q_n' \}$

Training Process:

7: Generate the set of N random hyperparameter combinations from *P*.

8: Create N different networks pertaining to *X* combinations found in *step 7*.

9: Train each pre-trained network PT_{net} on *X* from step 8.

10: Choose the most efficient PT_{nets} with specified accuracy threshold from step 9.

11: Perform grid search and assign suitable weights $w_{ts} = argmax \sum_{i=1}^{n} w_1, \ldots, w_n$ *to* PT_{net}.

12: Perform the final training for the best-M networks selected from *step 11* using the entire training dataset.

Testing Process:

13: Input testing image *q'* from the *Q'*

14: Generate output *h(q')* predictions from each of the chosen *PTnet* from *step 12*.

15: Perform the final classification by doing an ensemble of predictions from *step 14*.

Output: Final classification label *h(q')* for a testing image from *Q'*

DCGAN (Radford et al., 2015) is a convolutional network GAN model that improves the stability of the standard GAN model and the quality of generated images. By adding a transposed convolutional layer in the generator network of DCGAN, we create cDCGAN - a conditional deep convolutional GAN. The cDCGAN has a vital role in retaining and restoring the fine-grain details of a feature map. Due to these properties, cDCGAN is a suitable choice for our framework. This technique generates synthetic images of a specific class to handle class imbalance and introduce unseen data points, increasing the generalisability of the classifier.

We first train a neural network and a cDCGAN model on colonoscopy images to begin our approach. The cDCGAN generates new samples from a seed, a vector of random numbers. The seed determines the general appearance of the image based on the class label. In the first step, we define the initial generator parameters θ_G, discriminator parameters θ_D, and transposed convolutional parameters θ_T. The generator receives two inputs: a 100-node latent vector as the initial seed and class labels. Before passing through the rest of the network, the seed and class labels are concatenated and reshaped. The transposed convolutional layer then upscales the data to 28 x 28 pixels. Two distinct inputs are used for the discriminator in the cDCGAN model: a class label and an image generated by the generator. The class label is first converted into an embedding, a vector of learnt numerical values that represent the essence of the class label. This embedding is then passed through each layer of the cDCGAN along with the generated image. To create the final cDCGAN model, we merge the generator and discriminator models. The discriminator model is specified as non-trainable, as we trained it separately with real and synthetic data. In the next step, we train and evaluate the model. We then allow the model to generate images for each class label and store the synthetic endoscopic images

according to the individual class. Lines 1 - 7 of Algorithm 2 present the procedure.

Algorithm 2 The Training Scheme of Concordant cDCGAN and SIGB

Require: *X:* Training endoscopic image data;*Y:* Labels, where $Y = \{y \mid y \in \{0, 1, 2\}\}$; θ_G*:* G's parameters; θ_D*:* D's parameters; θ_T*:* T's parameters; *L:* Latent Space.

Input: *X*, where $\mathbf{x} \in \mathbf{X}$

1: Initialize parameters θ_G, θ_D, θ_T, L=100, Adam Parameters: α=2x10^{-3}, β1=0.9.
2: Input seed drawn from the latent space L and class labels to the generator G.
3: Upscale the image dimensions through transposed convolutional layers T and obtain the output image.
4: Input the image generated in *Step 3* and the image label to the discriminator D.
5: Combine G and D to create *cDCGAN*. Set model D as non-trainable to allow separate training.
6: Train the final model and generate conditional synthetic images *CSI*.
7: Store CSI_s in the relevant folder $F_{synthetic}$, where $F_{synthetic} = \{0, 1, 2\}$
8: Train CNN to determine concordance C
9: Input *CSI* to concordance classifier
10: **for** $n = 1 \rightarrow$ N **do** // where N is the total no. of components
11: **if** $Prob \geq 0.75$ **then**
12: store image CSI_n in corresponding folder F_{org} of original dataset.
13: **else**
14: discard image
15: **end if**
16: **end for**

The subsequent step in our methodology involves categorizing the classifier's

predictions into three distinct categories: strong concordance, weak concordance, and no concordance. This categorization is crucial for refining the quality of the synthetic images included in the dataset and for ensuring the robustness of the classification model. To determine the level of concordance, synthetic images are systematically extracted from the dataset and individually evaluated by the classifier. For each image, the classifier generates a prediction along with an associated probability score. If the classifier's probability score for a particular class is greater than or equal to 0.75, the prediction is classified as strong; conversely, if the score is less than 0.75, the prediction is deemed weak.

The categorisation process is further refined by comparing the classifier's prediction with the class label generated by the Conditional Deep Convolutional Generative Adversarial Network (cDCGAN). If a prediction is strong and aligns with the cDCGAN's class label, the image is considered to exhibit strong concordance, indicating a high level of agreement between the classifier and the generated label. This agreement suggests that the synthetic image is of high quality and closely resembles the features and characteristics of real images within that class. On the other hand, if the classifier's prediction is weak but still matches the cDCGAN label, the image is categorised as having weak concordance. This weak concordance indicates some level of uncertainty or less confidence in the classifier's prediction, even though it aligns with the synthetic label. Finally, images that do not match the cDCGAN class label or have a probability score below the threshold for strong or weak predictions are classified as having no concordance. These images are typically considered low quality or less reliable for training purposes.

The comprehensive overview of this methodology, including the classification

and concordance determination process, is visually represented in Figure 3.2. The methodology is designed to ensure that only high-quality synthetic endoscopic images, which demonstrate either strong or weak concordance, are included in the dataset for further analysis and model training. Images that fall into the category of low or no concordance are excluded from the dataset to prevent them from negatively impacting the model's learning process, as detailed in lines 8 to 16 of Algorithm 2. This selective inclusion and exclusion process helps to maintain the integrity of the training data and enhances the model's ability to generalise well to test data.

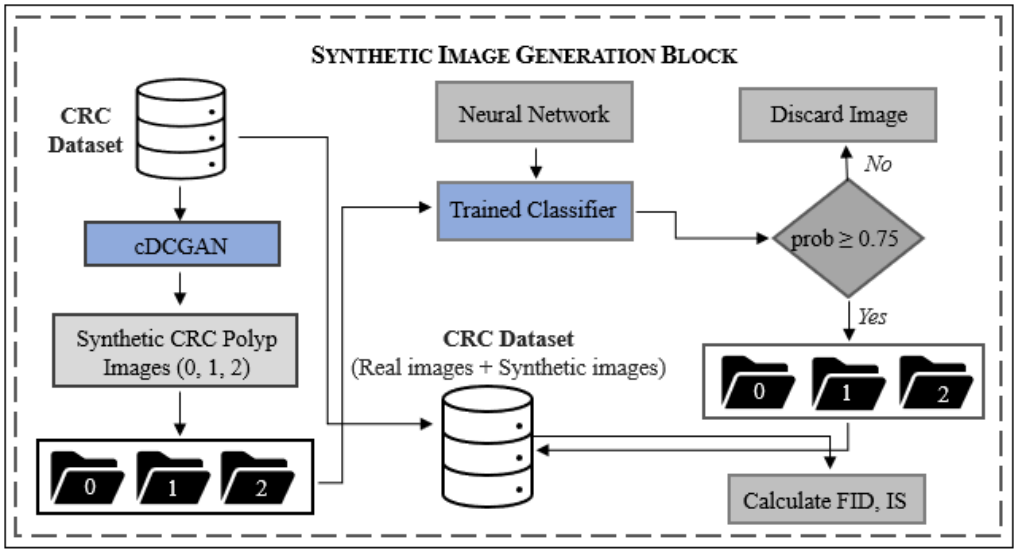

Figure 3.2: Framework of synthetic image generation block

To quantitatively ensure the high quality of the synthetic images that are being integrated into the dataset, we utilised the Fréchet Inception Distance (FID) metric. The FID score measures the distance between the distributions of real and synthetic images in a feature space, with lower scores indicating closer resemblance and higher quality. Inception Score (IS), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index

(SSIM) is used to assess the quality of synthetic images. By employing these metrics, we ensured that only synthetic images with acceptable values were included in the final dataset.

The output of the SIGB (Synthetic Image Generation Block), which comprises these high-quality synthetic images, is then merged with the existing dataset to create a more comprehensive and enriched dataset for subsequent training and evaluation phases. This strategic integration of high-quality synthetic data is intended to improve the robustness and accuracy of the classifier, ultimately enhancing the overall performance of the CAD system for the classification of colorectal polyps.

To illustrate the second phase of our methodology, we applied Algorithm 2 on images from dataset D to generate a synthetic image using cDCGAN I with the class label Adenoma. After generating an image, we proceed to determine whether it meets the criteria for inclusion in the current dataset. For that, three possible scenarios that could arise are given in Table 3.1. The first requirement is that the predicted and actual class labels must match. If the predicted class does not match the actual class label (Scenario 3 of Table 3.1), that image will be discarded. On the other hand, if the predicted label matches the class label, we will verify the class probability (Scenario 1 and 2 of Table 3.1). In case the probability is greater than or equal to 0.75, it will be categorized as highly concordant and added to the dataset; otherwise, it will be discarded.

Table 3.1: Concordance determination of generated images

Scenario	**Predicted Class**	**Predicted Class Probability**	**Concordance Category**	**Addition to Existing Dataset**
S1.	Adenoma	≥ 0.75	High	Yes
S2.	Adenoma	< 0.75	Low	No
S3.	Adenocarcinoma or Hyperplasic	< 0.75	No Concordance	No

3.1.3 Deep Feature Extraction Block (DFEB)

After generating a diverse and comprehensive dataset, the next crucial phase in our methodology involves extracting meaningful image features to address **RQ2**. To achieve this, we have utilized the ViT model, which has been specifically chosen for its advanced capabilities in deep feature extraction. The ViT model is highly effective in learning rich and nuanced representations from visual data, thanks to its innovative architecture that excels at capturing both global and local image patterns. Unlike traditional CNNs, ViT operates on image patches, which allows it to maintain high resolution and detailed representation of the image, making it particularly suitable for tasks requiring intricate feature extraction (Raghu et al., 2021).

The Vision Transformer leverages a powerful attention mechanism, which is central to its ability to focus on the most relevant parts of an image. This attention mechanism allows the model to identify and emphasize important regions of the image, thereby capturing subtle and intricate details that might be overlooked by other models. By attending to various parts of the image and understanding the context within those regions, ViT models can produce highly detailed and informative feature representations. This capability is especially beneficial in medical imaging applications, where the ability to discern fine details can significantly impact diagnostic accuracy. Incorporating the high-quality synthetic images generated in the preceding phase further enhances the effectiveness of the feature extraction process. By expanding the dataset with these synthetic images, the model is exposed to a broader range of variations and features, which enriches its learning experience and improves its generalization capabilities. The increased dataset size and diversity contribute to a more robust feature extraction process, enabling the Vision Transformer to capture a more comprehensive set of features from

the images.

Furthermore, ViT models are known for their computational efficiency when handling large datasets, which provides a distinct advantage over traditional CNN architectures. While CNNs are effective, they often require extensive computational resources and may struggle with large-scale datasets due to their inherent design. In contrast, the transformer-based approach of ViT allows for more efficient processing and scalability, making it well-suited for working with extensive and diverse datasets. The ability of ViT to handle large volumes of data while maintaining high performance in feature extraction is a key reason for its integration into our methodology. By leveraging the Vision Transformer for feature extraction, we ensure that our model benefits from the most advanced and effective techniques available. This approach allows us to capture detailed and accurate features from both real and synthetic images, thereby enhancing the overall performance of our classification system. The combination of high-quality synthetic images and the sophisticated feature extraction capabilities of ViT positions our methodology to achieve optimal results in addressing **RQ2** and advancing the field of automated colorectal polyp classification.

The ViT operates using the attention mechanism by establishing connections among neighbouring and distant pixels. To achieve this, the input image is partitioned into small patches, resulting in a 4-dimensional matrix organized by batch, row, column, and depth. To make the model compatible, we reshaped the original image x having the original shape [H × W × C] is restructured to $x_p \in \mathbb{R}^{N \times (P^2C)}$, here (H, W) represents image resolution, (P, P) represents the spatial dimensions of each patch, and C denotes the channel count (Gu et al., 2023). The number of patches N is calculated as given in

equations 3.1 - 3.2 below:

$$N = \frac{H \times W}{P^2} \tag{3.1}$$

where the patch size is represented by *P*. The image size is (244 x 244), and the patch size is (32 x 32); the calculation for the number of patches is represented as follows:

$$N = \frac{224 \times 224}{(32)^2} = 49 \tag{3.2}$$

A transformer pipeline was set up to normalize the images and reduce the influence of any biases. After that, the image *(x)* is converted into a flattened 2-D matrix M_{2D} as given in equation 3.3 and is projected linearly into 1-D embedding vector, M_{1D} as a 1-D token embedding sequence is the appropriate input format for a standard transformer as shown in equation 3.4.

$$M_{2D} = \begin{bmatrix} [x_1] & [x_2] & \dots & [x_7] \\ [x_8] & [x_9] & \dots & [x_{14}] \\ \dots & \dots & \dots & \dots \\ [x_{43}] & [x_{44}] & \dots & [x_{49}] \end{bmatrix} \tag{3.3}$$

$$M_{1D} = \begin{bmatrix} [x_1^1 & x_1^2 & \dots & x_1^{768}] \\ [x_1^2] & [x_2^2] & \dots & [x_2^{768}] \\ \dots & \dots & \dots & \dots \\ [x_1^{768}] & [x_2^{44}] & \dots & [x_{49}^{768}] \end{bmatrix} \tag{3.4}$$

Given the computational expense associated with transformers, particularly when applied to high-resolution images, the Vision Transformer (ViT) utilises a strategy in which image patches are aggregated into smaller segments. This aggregation process involves dividing the input image into a series of non-overlapping patches, each treated as a single input token. Subsequently, these patches are embedded using a positional embedding technique, allowing the model to maintain spatial relationships within the image while processing larger image sizes (Trockman & Kolter, 2022). Positional embedding, denoted as E_{POS}, plays a significant role in providing the transformer with information about the order and position of patches in the original image layout. The transformer model employs a combination of sine and cosine functions at varying frequencies to create positional embeddings, as proposed in the original transformer architecture (Vaswani et al., 2017). This sinusoidal approach to positional encoding is designed to allow the model to capture and understand the relative and absolute positions of patches within an image. Specifically, patches located at odd-numbered positions within the sequence are embedded using the cosine function, whereas patches positioned at even-numbered positions are embedded using the sine function. This alternating pattern of sine and cosine functions provides a unique encoding for each position, which helps the transformer to maintain a sense of order and location for each patch across different network layers.

In the context of positional encoding, *pos* represents the specific position of a patch within the sequence, while I denotes the dimension of the embedding space. Therefore, full positional embedding involves computing the sinusoidal values for each possible position across multiple dimensions. This encoding process ensures that each patch has a distinct positional representation that the model can use to learn spatial

dependencies and relationships between different patches. The maximum length of the patch group is represented by d, which corresponds to the number of patches that the transformer can handle at once. After the patches are linearly projected into a higher-dimensional space, each projected patch is augmented with its corresponding positional embedding. This augmentation process effectively combines the positional information with the patch's feature representation, allowing the transformer to not only recognise the features within each patch but also understand their spatial context relative to the entire image. This capability is particularly advantageous in medical image analysis, where the spatial arrangement of features can provide critical insight into the diagnosis and characterisation of various conditions, such as identifying different types of colorectal polyps. By embedding patches, the Vision Transformer can efficiently handle larger images and capture intricate details without losing important spatial information. This approach ensures that the model can effectively learn complex visual patterns and relationships within high-dimensional data, which is essential to achieve accurate and reliable performance in automated colorectal polyp classification. The combination of patch aggregation, positional embedding, and the transformative power of the attention mechanism positions the Vision Transformer as a powerful tool for deep feature extraction in advanced medical imaging applications.

Encoding an image involves several steps, including linear projection and positional embedding. Once these steps are completed, the embedded patch is then passed to the encoded block for further processing. The encoder comprises eight identical blocks, each comprising six layers. These layers include a normalisation layer, followed sequentially by a multi-head self-attention (MSA) layer and a multi-layer perceptron (MLP). This combination of layers helps to ensure that the encoded image is accurately

represented and can be used for further analysis or processing. In the encoder block, the input *(EP)* is combined with the output produced by MSA. The resulting output subsequently passes through a normalisation layer, followed by an MLP dense dropout layer. Lastly, a skip connection is established from the input and concatenated with the attention mechanism output. This enhances the positional impact by providing the subsequent layer with the original embedded patch and the concatenated attention output. The attention mechanism involves three embedding matrices: key K, query Q, and value V. These matrices are computed using weight matrices W_Q, W_K, and W_V, respectively, with the equations 3.5 - 3.7 (Dutta et al., 2023), where the embedded patch are represented by *EP*, and W_Q, W_K, $W_V \in \mathbb{R}^{d_{model} \times d_k}$ are the weight matrices.

$$Query, Q = EP.W_Q \tag{3.5}$$

$$Key, K = EP.W_K \tag{3.6}$$

$$Value, V = EP.W_V \tag{3.7}$$

The MHA layer executes the following equation in parallel multiple times to perform the single attention function called the head. The method for calculating attention is given in equation 3.8. Where d_k is a dot product that scales to prevent the explosion of the attention value. The attention value is a scoring function that represents the correlation between two image patches. The proposed framework includes multi-head attention *(MHA)* with twelve heads, which is represented in equation 3.9:

$$Attention, (Q, K, V) = softmax\left(\frac{QK^T}{d_k}\right)V \tag{3.8}$$

$$MHA = Attention, (Q, K, V)_{\text{x12}} \tag{3.9}$$

In the subsequent layers, a dense layer with Gaussian Error Linear Unit (GELU) activation is utilised in the multilayer perceptron, introducing nonlinear characteristics. Here, ϕ denotes the cumulative distribution function of the Gaussian distribution used in the GELU activation function. Finally, the features of the last layer of the model are extracted, which is $T_{feature\_embeddings}$ and the feature embeddings are flattened and placed in a single tensor using a tensor background using equations 3.10 - 3.12 (Dutta et al., 2023). All the embeddings obtained by the feature extractor are then stored in E_T for further framework modules. The learning scheme for this block is given in Algorithm 3, and the overview of the methodology is shown in Figure 3.3 and the incorporation of the ViT network into the network pipeline is shown in Figure 3.4.

$$GELU(x) = xP(X \geq x) = x\phi(x) \tag{3.10}$$

$$T_{\text{feature\_embeddings}} = GELU(MHA) \tag{3.11}$$

$$E_T = flatten\left(T_{\text{feature\_embeddings}}\right) \tag{3.12}$$

Algorithm 3 The Learning Scheme for DFEB

Require: $X_{(s,\,o)}$: Synthetic (s) and Original (o) endoscopy images, $Y_{(s,\,o)} = \{y \mid y \in \{\,0, 1, 2\,\}\}$

Input: Endoscopy Images, where $\mathbf{x} \in \mathbf{X_{(x,\,o)}}$

1: Load the image dataset D, where D contains $(X_{(s,\,o)})$, $(Y_{(s,\,o)})$.
2: Train Vision Transformer Model T on $(X_{(s,\,o)})$ with optimized network hyperparameters through grid search.
3: Divide the input image x into small patches of a 4D matrix.
4: Reshape the original image x into $x \in \mathbb{R}^{N\ x\ (P^2C)}$
5: Calculate the number of Patches N.
6: Define transformer pipeline P, that is, image resizing and data normalization.
7: Convert x_p into flattened 2D matrix M_{2D}.
8: Project linearly into 1D embedding vector M_{1D}.
9: Group image patches into smaller groups.
10: Embed image patches with positional embedding E_{POS}.
11: **if** POS/2 = 0 **then**
12: embed using *cosine* function
13: **else**
14: embed using *sin* function
15: **end if**
16: Concatenate linear projection patches with E_{POS} to produce embedded patches.
17: Pass embedded patches to encoded blocks.
18: Extract the feature embeddings E from the last layer of the trained transformer model for the model explainability block.
19: Convert the embeddings into a single tensor E_T and store resized tensor embeddings in a list. Convert E_T into matrix M for Manifold Learning Block

Output: Feature embeddings E_T, Feature matrix M

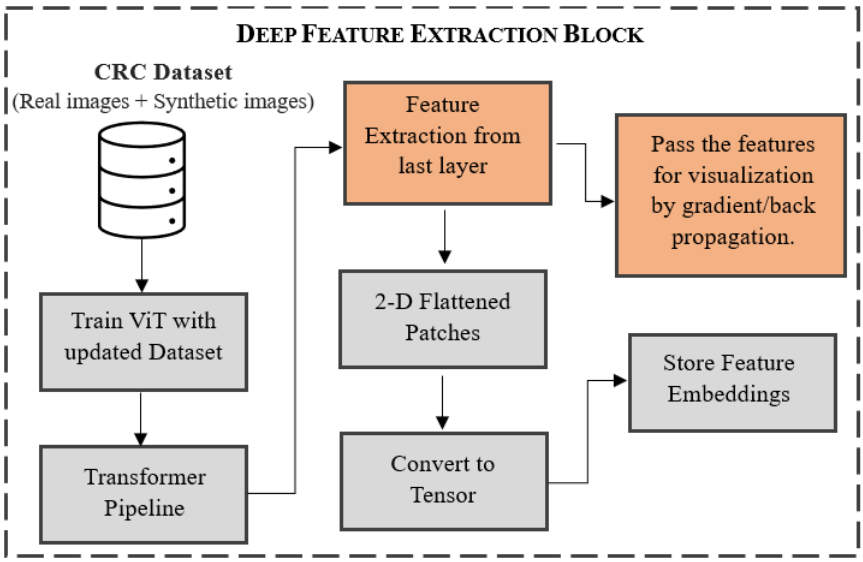

Figure 3.3: Methodology of deep feature extraction block

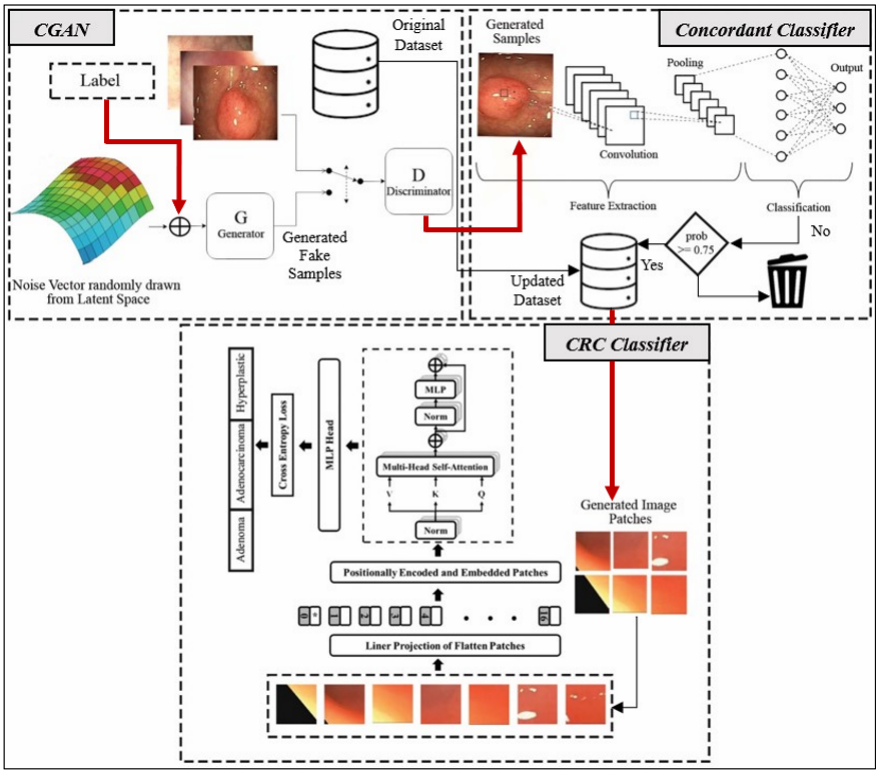

Figure 3.4: Feature Extraction using Vision Transformer

3.1.4 Model Explainable Block (MEB)

This step aims to provide further information about the decision-making process of our model addressing **RQ2a**, thus improving trust in the decision of the model. It explains the human level, which is crucial in the regulated healthcare sector, as it plays a pivotal role in bolstering confidence in the automated diagnosis system among specialists, patients, and other stakeholders (Nazir et al., 2023). Due to the opaque nature of DNNs, it is impossible to get insight into the features that contribute to the predictions. Therefore, the features obtained through the last layer of DFEB are utilised by this framework component to reveal the features chosen for the decision-making process. We incorporated visualisation using the gradient/backpropagation method for this purpose and used heat maps to present the hidden information regarding the importance of features. This approach utilises the backward pass to determine the contributing features and assess the region's significance for the model's prediction. We implemented Grad-CAM to visualise the contributing image regions instead of its recent versions due to its established robustness and reliability in several studies. Grad-CAM is a reliable approach for ensuring explainability, and the recent versions have higher computational overhead, so they restrict the processing where resources are constrained. Therefore, its versatility, lower computational cost and reduced complexity make it a suitable choice for our study.

In the first step, we determine the gradient of y_c using feature map activation A_k from the convolutional layers. The gradient in equation 3.13 is equal to the input image, representing the feature map.

$$Gradient = \frac{\partial y^c}{\partial A^k} \tag{3.13}$$

In the next step, we compute the alpha value. This can be achieved by calculating the average set of global variables against the breadth measure I and the height dimension index j. The result of the following calculation will give us α_k^c, the weights of neuron importance as shown in equation 3.14.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k} \tag{3.14}$$

In the final step, we combine the feature map activations A_k using a weighted approach; here, the weights α_k^c are added for the final calculation. The alpha value is used to calculate a weighted sum of feature maps. This sum is then passed through a ReLU operation, which zeros out negative values and keeps only positive values, resulting in the final Grad-CAM heat map as given in equation 3.15 (Mukhtorov et al., 2023).

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{3.15}$$

The highlighted regions of the image represent the features used for the decision-making process; these features are stored separately for further analysis. The learning scheme of this block is given in Algorithm 4.

3.1.5 Manifold Regularization Block (MRB)

To construct an accurate classifier, it is critical to create suitable feature vectors that can be effectively separated into different classes. To address **RQ3**, the features extracted through ViT in the previous step contain high dimensionality, and diminishing the curse of dimensionality is a crucial step to construct an efficient image classifier. Moreover,

performing regularization through manifold learning tackles the overfitting problem. Applying manifold regularization at the junction of feature extraction and final classification significantly improves the quality of feature representations for subsequent image classification. Since the obtained features are of the utmost importance in contributing to the decision of the model, we extract a subset of these features to reduce computational cost and regularise the model simultaneously. Therefore, we reduce the data dimensions in this stage by preserving statistical and geometric properties in lower dimensions.

Algorithm 4 Learning Scheme of Explainable Model with Grad-CAM

Input: Feature embeddings E_T

1: Compute g_c, the gradient, using A_k, the feature map activation, equation 3.13.
2: Compute α, equation 3.14.
3: Compute the average of global variables against i and j.
4: Compute a weighted combination of A_k by adding α_k^c to final calculation.
5: Apply ReLU operation on the computed weighted sum WS_{A_k}, equation 3.15.
6: **if** $WS_{A_k} \geq 0$ **then**
7: store WS_{A_k}
8: **else**
9: $WS_{A_k} = 0$
10: **end if**
11: Generate Grad-CAM heat map on obtained gradients.

Output: Heatmap of input $\boldsymbol{x}$

Various nonlinear dimensionality reduction algorithms were applied one by one to reduce the features and evaluate their effect on the concluding classifier, as shown in Figure 3.5 and passed on to the classifier for predicting the classes. Given the data matrix

M obtained from the feature extraction block, $M \in \mathbb{R}^{n\,x\,d}$, where $\mathbb{R}$ is the number of neighbors to consider for each point when reconstructing the local neighborhood, d is the desired dimensionality of lower dimensional space, and n represents the column of the matrix in $\mathbb{R}^{d}$. Using k nearest neighbours, every sample M_i can be linearly reconstructed $\hat{M}_i$ as shown in equation 3.16 (Samad & Sekmen, 2021).

$$\hat{M}_{\mathrm{i}} = \sum_{j} W_{ij} X_j, j = 1, ..., k \tag{3.16}$$

For each sample represented by index i, there are k neighbours identified through index j. In the equation 3.16, W is a square matrix of size n x n, serving as a weighted matrix with columns being the linear coefficients for neighbouring samples. These coefficients are employed to reconstruct rows from the individual samples. Zero is assigned as the weight of all nonlinear neighbouring samples, and the remaining weights are adjusted to fine-tune the W matrix through a least squares approach as presented in equation 3.17.

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{n} \|M_i - \hat{M}_i\|_2^2 \tag{3.17}$$

Therefore, W functions as the local map retained in a low-dimensional embedding. During the last phase, d-dimensional coordinates are derived as the columns of the U matrix, also known as low-dimensional embedding vectors. This process ensures the preservation of the acquired W matrix by fine-tuning the specified embedding cost in equation 3.18.

$$\underset{U}{minimize} = \sum_{i=1}^{n} \|U_i - \sum_{j} W_{ij} U_j\|_2^2 \tag{3.18}$$

Mitigation involves resolving the sparse $n \times n$ *W* matrix through Eigen decomposition. This process results in the *U* matrix containing the targeted eigenvectors or d-dimensions. The *m*-dimensional representation of *d*-dimensional data is achieved by mapping the data onto the initial *m (where d > m)* Eigenvector as shown in Equation 3.18. Finding the optimal number of dimensions for an accurate classifier is important. In contrast to linear-dimensionality techniques, determining the appropriate number of dimensions in non-linear techniques is challenging. The SVM is known to generalise well in high-dimensional spaces (Yan, 2019). Therefore, we trained an SVM classifier on reduced-dimensional data representations generated by several manifold algorithms to determine the optimal dimensions in the designed framework. The classification report was evaluated for the different number of components to establish the ideal number by assessing the accuracy and stability of the classifier's performance.

3.1.6 Concluding DNN Classifier

As illustrated in Figure 3.5, the finalised set of components of the previous step is then implemented to perform classification. We have integrated only two fully connected hidden convolutional layers to fulfil this purpose **RQ4**. The selected optimisation algorithm is SGDM, 100 number of epochs and a 0.001 learning rate as established by our study (Younas et al., 2023). We perform batch normalisation for every training batch as it tends to normalise activation functions in the middle layers of neural networks. The normalisation of the final layer of the deep network is the primary significance of batch normalisation (Bjorck et al., 2018). The fully connected (FC) hidden layer maps the dimensional feature vector *n_vflatten* obtained from the manifold algorithm into n_1 feature space y_{FC}. The last fully connected layer maps n_1 to the output layer with n_2

classification nodes. The manifold loss, computed batch-wise, is determined based on the response of the first hidden layer, as illustrated in equation 3.19.

$$\xi = \frac{1}{N}\sum_{i}^{N} ||y_{LLE}^{i} - y_{FC}^{i}||_2^2 \tag{3.19}$$

In this context, y_{FC} denotes the output of the first fully connected hidden layer, N signifies the batch size, and y_{LLE} represents local linear embedding (LLE) of the final convolutional response. Reducing the mean-squared error term during training encourages manifold embedding in the fully connected layer alongside the convolutional filter parameter optimisation. The final training goal is to minimise the cost function using equation 3.20 for every batch of training samples.

$$\underset{w}{\operatorname{argmin}} \left(\alpha * \sum_{i} (y_i - \hat{y}_i)^2 + \beta * \xi + \gamma * ||l||_2 \right) \tag{3.20}$$

$$L(\hat{y}, y) = -\sum y^{(k)} \log \hat{y}^{(k)} \tag{3.21}$$

Where, $(y_i - \hat{y}_i)^2$ is the squared classification error. l_2 represents the norm encompassing all weights, including both the filter parameters and the weights of fully connected layers. α, β, γ are positive tuning parameters that control the impact of classification error contributions, manifold regularization, and weight penalty terms in equation 3.20. The learning schemes of MRB and the final classifier are given in Algorithm 5. We utilized the multiclass cross-entropy loss function as given in equation 3.21 to calculate the loss between predicted y and ground truth ŷ labels. Here, $y^{(k)}$ is 0 or 1, indicating whether predicted label k is the correct classification (Zafar et al., 2018). The complete

methodological framework is presented in Figure 3.5.

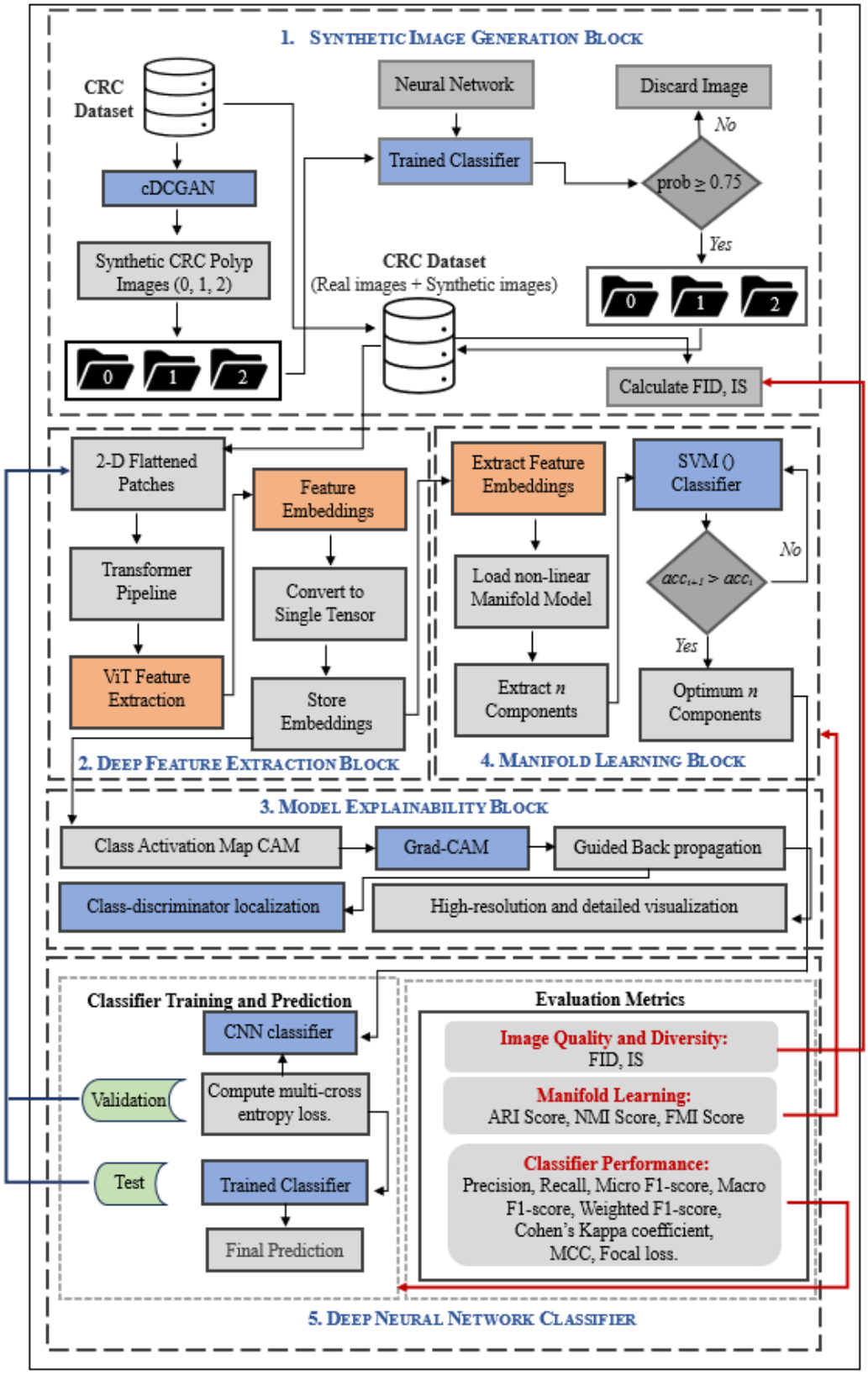

Figure 3.5: Methodological framework for CRC classification

Algorithm 5 Learning Scheme of Manifold Regularization and Concluding Classifier

Require: **M**: Feature Matrix

Input: Endoscopy Images, where $\mathbf{x} \in \mathbf{X_{test}}$

1: Load the nonlinear manifold learning model ML_{model}.
2: Build an SVM classifier with grid search for network optimization.
3: **Set** the parameters θ_{M} for ML_{model}, $i = 1$, $acc = 0$
4: **for** $q = 1 \rightarrow$ Q **do**, //where Q is the maximum number of components.
5: **Set** $d = q$
6: Load M from DFEB, //where $M \in \mathbb{R}^{\mathrm{n\ x\ d}}$
7: Linearly reconstruct M_i by apply *k-NN* algorithm to generate $\hat{M}_{\mathrm{i}}$, equation 3.16.
8: Reconstruct individual sample.
9: Assign zero weight to all non-neighbouring samples.
10: Update remaining samples weights to yield optimal W matrix, equation 3.17
11: Preserve the W matrix by optimize the embedding cost Q
12: Obtain d-dimensional coordinates as the column of U matrix
13: Train SVM() classifier on d manifold components
14: **if** $acc_{q+1} \geq acc_q$ **then**
15: q++, **goto** Step 4
16: **else**
17: d = q, *// where β is the input parameter to the classifier*
18: Initialize: Convolutional parameters, $\mathbf{P} = \{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_{\mathrm{n}}\}$
19: Initialize: Fully connected layer parameters, $\mathbf{F} = \{\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_{\mathrm{n}}\}$
20: **for** epoch = 1 $\rightarrow$ n_epochs **do**
21: Train DNN for d, P and F
22: Compute loss, equation 3.21
23: **end for**
24: Finalize model with calculated parameters
25: **end if**
26: **end for**

Output: Trained model predicts the corresponding class for input x

3.2 Summary of Chapter

This chapter comprises the methodology, which has been elaborately detailed with the aim of enhancing the functioning of computer-aided diagnosis (CAD) systems, with particular reference to the classification of colon polyps. The proposed conceptual framework provided illustrated each of its elements and assisted as ever in the understanding of the processes, the techniques used or the developments that ensued. The proposed methodology differs from other methods since it elucidates the use of sophisticated and advanced techniques such as data augmentation, feature extraction, and explainable artificial intelligence (XAI), among others. Moreover, it adds this extra nonlinearity through a derivative model restraining and polyp images reconstruction or modeling techniques which enhance image classification accuracy and reliability. Through this comprehensive method, the efficiency of the CAD system is increased, and the classification results are understandable and accurate, which is very important for the users of the systems in medical fields especially.

An integral component of the methodology we are proposing is using different types of data augmentation techniques, including the addition of synthetic concordant images. This method aims to overcome the limitations commonly known as the data bottleneck in the case of machine learning in most medical datasets. We have shown that these synthetic images are good even when used with transformers' attention mechanisms, a networking architecture class that emphasises important data features. Combining synthetic data generation and transformer models is essential in creating an interpretable AI system or model. This model serves to improve not only the performance of polyp imaging but also the performance of models that highlight the salient enablers of the final output of the model. These models are crucial in understanding the inner workings

of the model by allowing researchers and clinicians to see and understand the reasoning behind the prediction. Such point of view is very essential to makes AI transparent and improves trust and acceptance of the application of AIs in the medical field where critical decisions are made due to impact on the patients' health.

We have taken into account an emphasizing explanation of the application of advanced methods concerning dimensionality reduction in the course of the developed framework. In this way, such methods are instrumental in the reduction of the model complexity because of the decrease in the data size that requires being trained. However, while performing this reduction, enough care is made to develop the techniques so that they still give the model great ability to learn and encode the most important aspects of high-dimensional medical data. This is necessary in order to prevent degradation of the model performance and accuracy. These methods for dealing with high-dimensional data are especially helpful whenever combined with other model regularisation methods to deal with the overfitting problem which is a very serious challenge encountered when training deep learning models on datasets which are usually small as in the case of medical imaging data Overfitting constrains produces models that perform excellently on the training sample yet the performance on test samples is poor. These encourage effective nonlinear dimensionality reduction techniques and regularisation to be applied in a manner that prevents the model from overfitting the training data by making it, learn useful features rather than just patterns present in the training data only. This technique improves the strength of the model and guarantees its efficacy and consistency at making predictions under conditions of fewer or less diverse datasets. Through careful trade-off between the reduction in the complexity and saturation of the features, there is the building of more generalised models without compromising on the accuracy and

reliability in clinical settings with different patient populations.

This chapter provides the presentation in detail of the many algorithms and techniques which are employed in the methodology and explains how every part of the outlined framework works together towards enhancing the classification accuracy of the small medical data sets. By presenting the detailed function and interaction of each of the algorithms, the framework is shown how the components can interact with one another in a way that improves performance in terms of classification in a situation where the amount of data is constrained. Instead, this strategy helps reduce dependence on excessive data gathering and annotation practices that are usually labor-intensive, time-consuming, and at times unattainable in the medical field as a result of privacy issues and lack of annotated datasets.

In addition, the methodology also makes sure that even as significant attributes are extracted from the data, the most information is being put to use that is useful for the decision-making process of the model so that very little information is wasted. This is particularly important in areas like medical imaging where small details can be predictive of important clinical endpoints. At the same time, the strategy handles nonlinear high dimensionality of medical imaging data which has rich in patterns and variations. The nonlinear dimensionality reduction methods are included in the framework to make sure the geometric relations in the data are preserved which is helpful in retaining the important spatial and structural information needed for classification. This type of moderation in the complexity of the data also helps improve how the derived inferences are explained as well as strengthening the model's adequacy and practicability in deployment as would be required in the real world for the efficient performance of diagnosis as time is of the essence.

In order to validate our assumptions and showcase the effectiveness of our suggested approach in practice, we shall utilise several real-world datasets which will be described in detail in Chapter 6. This chapter is expected to be one of the most important parts of our work since it will provide quantitative proof of how effective and robust our methodology is cut out to be in tackling the multifarious and complex task of classifying colorectal polyps. In these results, we would therefore intend to demonstrate the efficacy of our approach in engaging with and resolving the variability and complications that are typically associated with the use of real medical images and, more specifically, the imaging of patients. Such demonstrations will also further emphasize the possible applicability of the proposed method in clinical practice where the use of reliable, accurate, and easy to interpret AI based diagnostic systems is on the rise. If such challenges are addressed, our methodology helps to break the current barrier in the development of CAD systems for the classification of colorectal polyps and also encourages the wider acceptance of AI technology in medicine.

The advent of this integration is mainly focused towards achieving a change in clinical practices that is significant through better and quicker diagnosis. This means that we have taken a big leap in reconciling the two critical aspects of AI in medicine; improvements in theory and the implementation of such advances within medicine, which helps to visualise AI as a powerful technology for the healthcare sector by proving its easiness, efficiency, and reliability in solving complicated diagnostic problems. Consequently, this adds to the increasing evidence base of the different areas in which AI can be applied in medicine and proceeds to make progress towards designing smart and efficient as well as patient-orientated healthcare.

Chapter 4

Experiments on Real-world CRC Datasets

This chapter describes the experiments performed on the three real-world datasets. In the first section, we elaborate on the network architecture and experimental setup used to test the framework developed, followed by the data sets used to validate the methodology. The evaluation metrics used for analysing the performance of the model are described in the last section. We have implemented the full-fledged prototype in various experimental settings and have conducted an extensive experimental evaluation to analyse performance.

4.1 Datasets for Learning Model

The proof of concept of the proposed CAD-CRC learning model uses the following three datasets:

4.1.1 UCI Dataset

The first dataset Gastrointestinal Lesions in Regular Colonoscopy is publicly accessible. This dataset has been utilised by other CAD researchers as well (Mesejo et al., 2016). The dataset comprises 76 short colonoscopy videos clinicians recorded under varying lighting conditions. The dataset includes both White Light (WL) and Narrow Band Imaging (NBI) data. All experimental procedures utilised digital images extracted from colonoscopy videos captured under narrow-band imaging (NBI) lighting, which is an advanced optical technique that enhances the visualisation of vascular patterns in the colon, providing detailed information about lesion types during the procedure (Shaban et al., 2019).

4.1.2 PICCOLO Dataset

The second dataset called the PICCOLO dataset (PICCOLO RGB/NBI Image Collection, 2021) was obtained from the Hospital Universitario Basurto, Spain. The data set includes clinical metadata as well as annotated frames from colonoscopy videos (Zhang et al., 2016). The metadata information for the acquired data and the annotation procedure are described in the following.

- The metadata completed by the gastroenterologist includes the number of polyps of interest, the current polyp ID, polyp size in mm, Paris classification, NICE

classification, and preliminary diagnosis.

- Metadata completed by pathologists includes final diagnosis and histological classification.

The dataset consists of 3,433 WL and narrow-band imaging (NBI) images extracted from clinical colonoscopy procedure videos of human patients. It encompasses 76 different lesions detected in 46 patients. The data set was divided into three subsets: the training set with 2,203 images, the validation set with 897 images, and the test set with 333 images. Details about the frames included in each set are provided in Table 4.1. The dataset covers three distinct types of polyps: Adenoma, Hyperplasia, and Adenocarcinoma.

4.1.3 CPDC Dataset

The Colonoscopy Polyp Detection and Classification Dataset CPDC dataset (Li et al., 2021) integrates publicly available endoscopic datasets, including MICCAI 2017, CVC colon DB, GLRC dataset, and KUMC (Kansas Medical Center) dataset. It comprises two classes: Adenomatous and hyperplastic polyps. Our experimental setup uses 116 training, 17 validation and 22 test sequences, with 28,773, 4,254, and 4,872 frames allocated for each set. The details of these frames in each dataset are provided in Table 4.1, and sample images are given in Figure 4.1.

4.2 Experimental Setup

To successfully train a Conditional Generative Adversarial Network (cGAN) for medical imaging, substantial computational power is necessary. The adversarial nature of

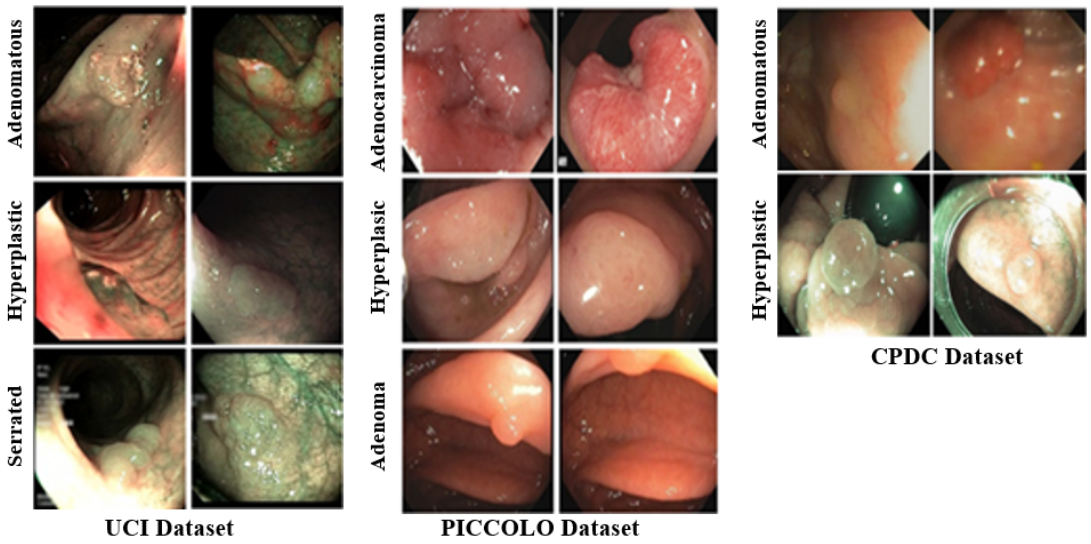

Figure 4.1: Sample polyp images of datasets

cGANs, where the generator and discriminator are trained simultaneously, demands high-performance GPUs or TPUs to handle complex computations efficiently. In addition, medical images, particularly high-resolution endoscopic or histopathological images, require significant memory, with a recommended VRAM of at least 16GB to manage large batches without bottlenecks. Training time is another key consideration: depending on the size of the dataset, the architecture of the model, and the tuning of the hyperparameters, the training can extend from several hours to multiple days, even on high-end GPUs.

All experimental procedures were performed on a GPU-based workstation featuring a 512GB hard drive, 32GB RAM, an Octa-core processor and a GeForce GTX 1080Ti GPU running on the Microsoft Windows 10 Operating System. The primary software used in the experimental study includes MATLAB and Anaconda Navigator. MATLAB was used in the preliminary study, while the Anaconda Navigator Jupyter Notebook was used mainly to develop the framework in Python programming language

4.2.1 Scalability Considerations for Clinical Deployment

Beyond computational requirements, the deployment of the trained model in a clinical setting presents additional challenges, particularly in scalability and efficiency:

- **Preprocessing Pipeline Optimisation:** Preprocessing raw medical images before feeding them into deep learning models significantly enhances computational efficiency. The workload on high-performance GPUs can be minimised using CPU/GPU optimisation, noise reduction, and data augmentation techniques. Additionally, training the model offline and storing synthetic samples in a structured database reduces real-time computational costs during inference.

- **Batch Inference for Throughput Enhancement:** Efficient batch processing of endoscopic images can improve overall model throughput, ensuring that real-time clinical applications are not delayed. Implementing optimized data pipelines that allow for batch inferencing can enhance model responsiveness.

- **Edge Deployment Considerations:** In resource-constrained environments such as real-time clinical workflows, deploying lightweight variants of GANs, such as MobileGAN or TinyGAN, can be beneficial. These models reduce computational overhead while maintaining performance, making them more suitable for deployment in endoscopy suites, mobile diagnostic tools, or embedded systems.

- **Parallelization & Distributed Training:** For large-scale medical image analysis, leveraging multi-GPU or distributed training strategies can accelerate model convergence. Frameworks such as TensorFlow's MirroredStrategy or PyTorch's Distributed Data Parallel (DDP) allow seamless parallelisation, ensuring that extensive datasets can be efficiently processed. This approach is particularly relevant

Table 4.1: Datasets details

Dataset	Category	Items	Training Set	Validation Set	Test Set
PICCOLO	Image type	WL	1,382	558	192
		NBI	821	340	141
	Diagnosis	Adenocarcinoma	172	166	127
		Adenoma	1,552	592	92
		Hyperplasia	435	139	114
		N/A	44	-	-
CPDC	Diagnosis	Adenomatous	18,241	2,977	3,410
		Hyperplastic	10,532	1,277	1,462

for institutions or research laboratories with large colonoscopy data sets that require high-throughput training and inference.

4.3 Network Architecture

The success of a deep learning model depends heavily on the quality and quantity of data used for training. Unfortunately, the collection of 76 colonoscopy images from the UCI Repository and 3433 images from the PICCOLO dataset is insufficient for our model. We need a much larger dataset to achieve accurate classification. Therefore, we have implemented an oversampling technique to enhance the dataset. With this approach, we can ensure that our model receives the best possible training data, resulting in accurate and reliable results. The training dataset was confirmed to be imbalanced, and label information was extracted. The data was split into an 80% training set and a 20% testing set. Variable labels were extracted from the training set, and group indexes were obtained for each class. The minority classes were oversampled in comparison to the majority class images. Due to the small and dissimilar dataset, developing an effective solution is challenging. If we train the model too deeply, it may overfit and become ineffective. Data augmentation is an indispensable tool for achieving optimal results in transfer learning.

Using this technique, we can effectively address problems and ensure a seamless data transition. Therefore, data augmentation was employed to facilitate successful transfer learning.

Deep learning architectures are particularly suited for classification tasks. However, pre-trained models can significantly simplify model development since they can be adjusted and customized for new tasks (Heinz et al., 2022). Furthermore, deep learning benefits greatly from transfer learning, where a model built for a specific task is used for a different task. In transfer learning, the model is trained on public datasets, and the initial weights of this model are used for the new task instead of assigning random weights, which is typically done when designing a network from scratch. In the regular training process, the last fully connected layer is responsible for final classification by presenting new images to the network and adjusting the weights of specific layers. The pre-trained model is fine-tuned, fully connected, and modified for the classification task. Each model is trained separately on oversampled and augmented images according to the specified training options for image classification. Our experimentation involved six CNN architectures: GoogLeNet, ResNet-50, Inception-v3, Xception, DenseNet-201 and SqueezeNet.

Improving the classification generalization through an ensemble-learning model can aid in dealing with the intricate characteristics of polyps. As part of our experimentation, various models were unified by integrating base learners for feature learning to build an ensemble learning module. In the initial training phase, three deep pre-trained CNNs with different architectural designs (GoogLeNet, Xception, ResNet-50) that perform residual learning were independently trained using ImageNet. In the final classification, an averaging-based ensemble learning approach is used where the averaging of

base-classifier weights is incorporated to make a final decision.

One of the goals of our framework is to generate high-quality synthetic images for model training; therefore, the data was up-sampled to a larger output feature map, which was done by embedding a transposed convolutional layer in GAN instead of a regular convolutional layer, which usually down-samples the data. The addition of a transposed convolutional layer in the generator network of cGAN creates cDCGAN - conditional deep convolutional GAN. cDCGAN plays a vital role in retaining and restoring the fine-grain details of a feature map. These properties of cDCGAN made it a suitable choice for our framework as this technique generates synthetic images of a specific class to handle class imbalance and introduces unseen data points for increasing the generalizability of the classifier. Our approach starts with training a neural network and cDCGAN model on colonoscopy images. cDCGAN generates unseen samples from a seed, the vector of random numbers; the seed determines how the image will generally look according to the class label. Initial generator parameters θ_{G}, discriminator parameters θ_{D}, and transposed convolutional parameters θ_{cT} are defined in the first step. Two inputs are passed to the generator: a 100-node latent vector as the seed and class labels. The seed and class labels are reshaped and concatenated before going through the rest of the network by the transposed convolutional layer that upscales the data to 28 x 28 pixels. Two separate inputs are also specified for the discriminator: a class label and an image generated by the generator model. The specified class label is converted into embedding and passed through each layer of cDCGAN, where the embedding is a vector of numbers learned during training that encodes the essence of the class labels. To create the final cDCGAN model, the generator and discriminator models are merged; here, the discriminator model is specified as non-trainable as we trained

the discriminator separately with real and synthetic data. In the next step, the model is trained and evaluated. Now, we allow the model to generate images for each class label, and the synthetic endoscopic images are stored according to the individual class.

In the next step, the predictions were categorised into strong, weak, and no classifier concordance. To verify the concordance of the classifier, synthetic images are extracted from the dataset individually, which is then predicted by the classifier. Predictions are considered strong if the classifier's probability is ≥ 0.75 and weak if the probability is < 0.75. If the predictions are strong and match the cDCGAN class label, the image is considered to have a strong concordance. On the other hand, if the classifier has a weak prediction even though it matches the cDCGAN label, it is considered to have a weak concordance .

Finally, we include the strongly concordant images in the dataset for further analysis and exclude those with low or no concordance. Therefore, this block generates only high-quality synthetic endoscopic images, part of the existing dataset. The outcome of the SIGB block is merged with the existing dataset to be utilised by further blocks of the proposed framework. After that, the deep feature extractor, a pre-trained Vision Transformer (ViT) model, acquires the initial data representations and extracts the features of concordant images. These models can learn rich representations from visual data by capturing global and local image patterns without degrading the image resolution and enabling effective feature extraction (Raghu et al., 2021). As ViT models are based on an attention mechanism, they allow the model to focus on relevant parts of images to capture the fine-grain details. The addition of strong, concordant synthetic images generated in the previous step transforms the data into a more extensive dataset. Compared to conventional CNN architectures, ViT models are computationally efficient

for large datasets. A standard input of a 1-D token embedding sequence was given to the transformer after reshaping the original image, x, with the original shape into flattened 2-D patches. A transformer pipeline was created where images were normalised to reduce the influence of any biases. Feature embeddings are generated using a tensor background and converted into a single tensor, and then all the embeddings obtained by the feature extractor are passed on to the next block.

Once the features were extracted effectively by the model, heat maps were constructed via Grad-CAM to incorporate XAI in the decision-making process. At the same time, it is critical to create suitable feature vectors that can be effectively separated into different classes. The features extracted in the previous step contain high dimensionality, and alleviating the curse of dimensionality is essential for constructing an efficient image classifier. Therefore, we reduce the data dimensions in this stage by preserving statistical and geometric properties in lower dimensions. Several nonlinear dimensionality reduction algorithms were applied to extracted features and passed on to the classifier to predict the classes. However, finding the optimal dimensions for an accurate classifier is crucial.

In contrast to linear dimensionality techniques, determining the appropriate dimensions in nonlinear techniques is challenging. Therefore, to assess the optimal number of dimensions in the proposed method, we trained an SVM classifier on reduced dimensional data representations generated by several manifold algorithms. The classification report was evaluated for the different number of components to establish the ideal number by assessing the accuracy and performance stabilization of the classifier. With a large number of images, model overfitting is evident, and nonlinear dimensionality reduction was used not only to handle the dimensions while preserving the geometry but also to

regularize the model. The finalized set of components from the previous stage is then deployed to perform classification. The components are fed into the final CNN classifier that predicts the class label. We utilised the multiclass cross-entropy loss function in equation 3.21 to calculate the loss between predicted y and ground truth $\hat{y}$ labels. Here, y(k) is 0 or 1, indicating whether predicted label k is the correct classification.

4.4 Evaluation Metrics

Accessing the performance of the developed model is an integral part of the study. As the framework is designed to address multiple issues, every component must be evaluated separately and thoroughly. The advanced evaluation measures used for each block are described below:

4.4.1 Classification Performance Analysis

The recall rate is a commonly used evaluation metric for classification models. The classification of colorectal polyps presents a class imbalance issue; therefore, individual and group models are evaluated using the F1 score. The F1 score is ideal for the evaluation of unbiased imbalanced datasets as it considers precision and recall equally.

The dataset contains various imbalances. When dealing with imbalanced data, it is important to use advanced metrics to evaluate the results. In addition, this project aims to classify the data into three distinct classes. Therefore, the proposed method was evaluated not only by accuracy, recall, precision, and F1-score but also by macro F1-score and weighted F1-score.

Two common methods for interpreting confusion matrices in multiclass settings

are micro F1-score and macro f1-score. Confusion matrix of every class $\mathbf{g_i} \in \mathbf{G} = \mathbf{\{\ 1, ..., K\ \}}$ such that the i-th matrix takes $\mathbf{g_i}$ class as the positive class and the rest of the classes $\mathbf{g_j}$ with $\mathbf{j \neq i}$ being negative classes. Micro-averages the performance over all samples to compute overall performance. The micro-averaged F1 score is calculated from the micro-averaged recall (R_{micro}) and micro-averaged precision (P_{micro}). This score is used to evaluate the overall performance of a classifier based on the weighted average of the F1-score for each class. The mathematical equations of these metrics are shown in the equations. 4.1, 4.2, and 4.3.

$$P_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} TP_i + FP_i} \tag{4.1}$$

$$R_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} TP_i + FN_i} \tag{4.2}$$

$$F1_{micro} = 2\frac{P_{micro} * R_{micro}}{P_{micro} + R_{micro}} \tag{4.3}$$

A high value of F1micro indicates that the model has performed well overall. The micro-average is misleading for imbalanced data as it fails to account for variations in predictive performance across classes. The macro-average calculates the average performance across all classes instead of evaluating each class individually. When the macro-value F1 is higher, individual classes perform well. Mathematical formulas are given in equations. (4.4), (4.5), and (4.8).

$$P_{macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FP_i} = \frac{\sum_{i=1}^{|G|} P_i}{|G|} \tag{4.4}$$

$$R_{macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FN_i} = \frac{\sum_{i=1}^{|G|} P_i}{|G|} \tag{4.5}$$

$$F1_{macro} = 2\frac{P_{macro} * R_{macro}}{P_{macro} + R_{macro}} \tag{4.6}$$

$$W = \frac{\sum_{i=1}^{|n|} w_i X_i}{\sum_{i=1}^{|n|} w_i} \tag{4.7}$$

$$L(\hat{y}, y) = \Sigma_k^k \, y_{(k)} \, log\hat{y}^{(k)} \tag{4.8}$$

4.4.2 Reliability Analysis

In imbalanced class problems, the Cohen's Kappa Coefficient is useful for evaluating performance and analyzing reliability. In equation 4.9, p_o is the aggregate accuracy, and p_e depicts the comparison between the model's prediction and the actual class value, which is determined by chance agreement. The coefficient results are interpreted as follows: values ≤ 0 indicate no agreement, values between 0.01 and 0.20 indicate none-to-slight agreement, values between 0.21 and 0.40 indicate fair agreement, values between 0.41 and 0.60 indicate moderate agreement, values between 0.61 and 0.80 indicate substantial agreement, and values between 0.81 and 1.0 indicate almost perfect agreement.(McHugh, 2012). The weighted average is shown in equation 4.7. For binary classification, the Matthew Correlation Coefficient (MCC) is considered, serving as a measure of the quality of binary classification, as presented in equation 4.10

$$kappa(k) = \frac{p_o - p_e}{1 - p_e} \tag{4.9}$$

$$MCC = \frac{TN \times TP - FN \times FP}{(TP + FP)(TP + FP)(TN + FN)} \tag{4.10}$$

4.4.3 Image Quality and Diversity Analysis

To assess the standard and diversity of generated images constructed by CRcGAN, four metrics namely: ***FID*** - Fréchet Inception Distance, ***IS*** - Inception Score, ***PSNR*** - Peak Signal-to-Noise Ratio and ***SSIM*** - Structural Similarity Index were used to ensure that images are of good quality. FID metric compares the feature representation of both image types by employing a pre-trained inception model. The mathematical equation to calculate FID is given in Equation 4.11, where μ_x and μ_y indicate the feature-wise average of synthetic and real samples, Σ_x and Σ_y are the covariance matrices of both images, and T_r represents the linear algebra operation trace. Lower FID values indicate good similarity in real and generated images, whereas FID less than ten shows that the generated samples are comparable to real samples.

$$FID = \| \mu_x - \mu_y \|^2 + T_r(\Sigma_x + \Sigma_y - 2\left((\Sigma_x\Sigma_y)^{1/2}\right) \tag{4.11}$$

Another metric to assess image quality is ***IS*** - Inception Score, the most widely used scoring algorithm for GANs. The mathematical equations are shown in equation4.12 where $(p(y|x))$ is the conditional probability, *p(y)* is the marginal probability distribution and D_{KL} shows the divergence between conditional and marginal probability. A higher value of IS indicates good diversity among the generated images.

$$IS(G) = \exp(E_x \approx p_g D_{KL}(p(y|x))p \parallel (y))) \tag{4.12}$$

PSNR is a widely used metric for assessing the quality of a reconstructed or compressed image/video compared to its original counterpart. It is particularly useful in image processing, medical imaging, and video compression to quantify the level of distortion introduced by compression or noise. The Peak Signal-to-Noise Ratio (PSNR) between an original image x and a processed image y is expressed in decibels (dB), which is calculated using equation 4.13, where MAX is the maximum possible pixel value and MSE is the mean square error, given by equation 4.14:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right) \tag{4.13}$$

$$MSE = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [x(i,j) - y(i,j)]^2 \tag{4.14}$$

SIGB generated multiple images that require the quality assessment of all synthetic images. Therefore, an average PSNR was calculated on multiple images. For a data set with N image pairs, the arithmetic mean PSNR is computed using equation 4.15:

$$MeanPSNR = \frac{1}{N} \sum_{i=1}^{N} \text{PSNR}_i \tag{4.15}$$

A higher PSNR value, ranging between 30 and 50 dB, indicates a high-quality reconstruction with less distortion. The lower PSNR between 20 and 30 dB represents a higher noise with noticeable distortion, but the images are still acceptable. A value of < 20 dB shows a significant loss of quality.

SSIM is a widely used metric to measure the perceptual similarity between two images, often used in medical imaging, image compression, and deep learning-based

reconstruction tasks. Unlike PSNR, which only considers pixel-wise differences, SSIM accounts for luminance, contrast, and structure, making it more aligned with human visual perception. The SSIM between an original image x and a processed image y is defined as shown in equation 4.16, where, μ_x, μ_y are the mean intensities of the images x and y, σ_x^2, σ_y^2 are the variances of x and y, σ_{xy} is the covariance between x and y and C_1 and C_2 are small constants to stabilise the division. SSIM values range from -1 to 1, where 1 indicates perfectly identical images, >0.9 shows a high similarity, and a value < 0.5 represents significant distortion.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.16}$$

The mean SSIM was calculated on multiple images in a dataset containing N image pairs, the mean SSIM is calculated using the equation 4.17, where $SSIM(x_i, y_i)$ is the SSIM score for each image pair (x_i, y_i).:

$$MeanSSIM = \frac{1}{N}\sum_{i=1}^{N} SSIM(x_i, y_i) \tag{4.17}$$

4.4.4 Manifold Learning Analysis

The performance of the multiple learning algorithms was evaluated on three different matrices: Adjusted Rank Index (ARI), Normalised Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). ARI is the measure of similarity between two clusters. It is calculated using the equation 4.18 (Rand, 1971; Hubert & Arabie, 1985; Vinh et al., 2009), where n is the total number of elements and $\binom{n}{2}$ is the total number of instances. The number of elements at the intersection of the clusters is represented by

n_{ij}, where *i* indicates the first cluster and *j* is the second cluster. The sum of elements in clusters *i* and *j* are represented by a_i and b_j. The index value ranges from -1 to +1, where a value of 0 indicates that there is a random agreement between clusters, -1 represents that the clusters are dissimilar, and +1 shows a perfect agreement between the clusters (Wagner & Wagner, 2007). Therefore, the higher index value indicates that the clusters are close to each other.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \tag{4.18}$$

NMI is a mutual information metric to determine the quality of the clustering using class labels. NMI is the normalised value of Mutual index (MI), presented in equation 4.19, where $|U|$ indicates the number of clusters in *U* clustering, whereas $|V|$ is the number of clusters in *V* clustering. *P(i)*, *P(j)* are the probabilities of elements *i* and *j* randomly chosen from cluster *U* and *V* respectively, and *P(i, j)* is the joint probability. NMI value ranges from 0 to 1, where 1 shows that the clusters are identical and 0 shows otherwise (Kvålseth, 2017). Equation 4.20 presents the formula for calculating NMI, where *H(U)*, *H(V)* are the entropy of clustering U and *V*.

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P(j)}\right) \tag{4.19}$$

$$NMI(U,V) = \frac{MI(U,V)}{mean(H(U),H(V)} \tag{4.20}$$

$$FMI = \sqrt{\frac{TP}{(TP + FP)(TP + FN)}} \tag{4.21}$$

FMI is an evaluation measure that determines the similarity between multiple clusters and measures the confusion matrices using equation 4.21. The formula is defined by the geometric mean of precision and recall of clustering. This measure handles the issue of varying cluster sizes and number of clusters. A higher value represents a higher similarity between clusters. The value of FMI ranges from 0 to 1, where 0 represents the worst or no agreement between the clusters, and 1 represents the best classification or perfect agreement between the clusters (Chicco & Jurman, 2023).

4.5 Summary of Chapter

This chapter provides a comprehensive summary of the experiments conducted throughout this thesis. Initially, we defined the core problem to address the challenges in CRC polyp image classification. To achieve this, we employed three extensive datasets: UCI, PICCOLO, and CPDC, which include both public and private sources. Sample images from these datasets are illustrated in Figure 4.1. The datasets were systematically divided into training, validation, and test sets to ensure a well-balanced representation of each cancer type within each subset. This approach aimed to maintain the integrity of the data while providing a reliable basis for model training and evaluation. Traditional data augmentation techniques, such as rotation, flipping, and scaling, were applied to enhance the dataset's diversity and volume. Additionally, we implemented a novel approach for synthetic image generation to further increase dataset variability and quality. The synthetic images produced were categorized based on their quality and diversity. This

categorization ensured that only high-quality synthetic images were incorporated into the training dataset, thereby enriching the dataset while avoiding potential degradation of model performance. Once the augmented dataset was prepared, it was reformatted to meet the dimensional requirements of the transformer model used for feature extraction. This transformation was crucial for enabling the model to process and analyze the data effectively. The process involved scaling and reshaping the data to fit the input specifications of the Vision Transformer (ViT), thereby optimizing the feature extraction phase and contributing to the overall robustness and generalization of the classification model.

In this research, a comprehensive examination and utilization of several advanced deep learning architectures were conducted, encompassing GoogLeNet, ResNet-50, Inception-v3, Xception, DenseNet-201, and SqueezeNet. Each of these models was rigorously customized with optimized hyperparameters, including the selection of the optimizer, learning rate, and number of epochs. These hyperparameters were meticulously fine-tuned using an extensive grid search methodology to ensure the ultimate performance of each model. To further enhance model generalization and robustness, ensemble learning techniques were applied. This approach involved aggregating the strengths of multiple weak base learners to create a more robust and generalised classification model. By leveraging the diversity and unique strengths of different models, ensemble learning significantly improves overall performance and reliability. Beyond the conventional deep learning models, the cDCGAN (Conditional Deep Convolutional Generative Adversarial Network) model was employed to train on colonoscopy images, generating synthetic concordant images. The goal of creating these synthetic images was to augment the dataset and address the critical issue of limited data availability. To effectively extract feature embeddings from the processed images, a transformer model was utilised. This

model used a tensor-based background to capture and represent complex features with high accuracy, thereby enhancing the overall feature extraction process and contributing to the robustness of the classification system.

To ascertain the performance of the model, thorough experiments were carried out, adopting various advanced evaluation metrics, including the performance of individual components of the framework. This examination of block by block enables the understanding of the contribution of each component to the overall performance, as this is done prior to going to the next steps in which all pieces are integrated in the model evaluation process. In view of the analysis of the image aesthetic aspects and its variability, we used such scoring metrics as FID, IS, PSNR and SSIM, which are the most effective metrics to measure the quality of images synthesized with GANs. These measurements are necessary to validate the quality of synthetic images generated through the cDCGAN model so that the data produced can be classified efficiently.

Given the imbalanced nature of the dataset, conducting a thorough reliability analysis was a crucial component of the evaluation method. To assess how well the classification model performed, particularly in the context of the imbalanced dataset, two key metrics were employed: the Cohen Kappa Coefficient and the Matthew Correlation Coefficient. These metrics are instrumental in quantifying the level of agreement between the predicted classes and the actual class labels. Specifically, the Cohen Kappa Coefficient measures the agreement between predicted and true class labels while accounting for the agreement that might occur by chance, providing a nuanced understanding of model accuracy. Similarly, the Matthew Correlation Coefficient offers a comprehensive measure of the model's performance by evaluating the quality of the binary classifications in terms of true positives, true negatives, false positives, and false negatives. By utilizing these

metrics, the analysis effectively captures how well the model performs on the imbalanced dataset, offering insights into its reliability and overall effectiveness in predicting class labels amidst uneven class distributions.

When evaluating the outcomes of applying nonlinear dimensionality reduction techniques, several metrics proved to be highly effective: the Adjusted Rand Index (ARI), the Fowlkes-Mallows Index (FMI), and the Normalised Mutual Information (NMI). These metrics are crucial for assessing and quantifying the success of the dimensionality reduction process as they help to verify that the reduced dimensions still preserve valuable structures and relationships within the data. The ARI measures the agreement between the clustering results and the ground truth, adjusting for chance, while the FMI evaluates the similarity between clusterings in terms of both false and true positive rates. NMI assesses the amount of information shared between the clusters and the true class labels, providing insight into how well the reduced dimensions represent the original data.

To enhance the model's ability to generalise and prevent overfitting, regularisation was achieved through the use of the multiclass cross-entropy loss function. This function is pivotal in training deep learning models to effectively manage complex, high-dimensional data. By minimising the cross-entropy loss, the model is encouraged to produce probabilistic outputs that are closer to the true class distributions, thereby improving its performance on unseen data.

The classification performance of the model was thoroughly assessed using a comprehensive suite of metrics, including sensitivity, specificity, accuracy, recall, precision, F1-score, and additional measures such as micro, macro, and weighted F1-scores. Sensitivity and specificity provide insights into the model's ability to correctly

identify positive and negative instances, respectively, while accuracy offers an overall measure of correctness. Recall and precision evaluate the model's performance in retrieving relevant instances and avoiding false positives. The F1-score, along with its micro, macro, and weighted variations, offers a balanced view of performance considering both precision and recall, providing a nuanced evaluation of the model's effectiveness across different classes.

To benchmark the efficacy of the proposed model, comparisons were conducted with existing studies that utilised the same dataset. This comparative analysis was instrumental in highlighting the strengths of the proposed model and demonstrating its superior performance in managing complex and high-dimensional data relative to previous approaches. The results underscored the improved ability of the model to capture intricate patterns and deliver more accurate classifications, marking a significant advancement in the field.

Chapter 5

Results and Discussion

This chapter analyses and demonstrates the experiment results from the designed framework and experiments performed on three real-world datasets. The performance of every module of the framework is analysed and discussed. The results are validated on the basis of the evaluation metrics specified in the previous chapter. Section 5.1 elaborates on the experimental results obtained by the ensemble learning block. Section 5.2 provides comprehensive results on synthetic image generation and deep feature extraction blocks. The model explainability and manifold regularisation results are discussed in section 5.3. Finally, the classification results are demonstrated in Section 5.4.

5.1 Ensemble Learning Block

This block demonstrates the results obtained by the experiments conducted on the designed ensemble learning block. First, we discuss the performance of transfer learning-based experimentation on benchmark and balanced datasets. Second, we elaborate on the results produced by the ensemble model, where we discuss the performance of the proposed classifier, provide a comparative analysis with existing studies on the same datasets, and analyse and discuss the other important evaluation measures.

5.1.1 Transfer Learning-Based Polyp Classification

The experiments were conducted on the UCI benchmark dataset using six modern CNN architectures: GoogLeNet, ResNet-50, Inception-v3, Xception, DenseNet-201, and SqueezeNet. The rationale behind selecting this particular set of architectures is their proven performance for classification tasks in literature. The experiments were designed to find the best hyperparameter settings for the neural networks for classification purposes that will aid in addressing the research question **RQ1a**. Furthermore, the most efficient network was also selected by the experiments to address the research question **RQ1b**.

E 1.1: Performance on Benchmark Data

The pre-trained network was applied to the training set to identify the optimal optimiser for the given dataset. The three optimisers selected were ADAM - Adaptive Moment Estimation, SGDM - Stochastic Gradient Descent, and RMSprop - Root Mean Square Propagation. Optimisation algorithms can reduce error rates in deep neural network models during training. The Adam optimiser is widely favoured for its minimal need for

tuning and demonstrated effectiveness in enhancing model performance. This study compared various optimisers to find the best fit with other hyperparameters. The experimental configuration employed for this section is provided in Table 5.1.

Table 5.1: Experimental settings

Networks	optimisers	Learning Rates	Epochs
GoogLeNet	ADAM	0.001	50
ResNet-50			
Inception-v3	SGDM		100
Xception			50
DenseNet-201	RMSprop	0.005	
SqueezeNet			100

Since our study is conducted on balanced data, and our dataset is also balanced, we consider the SGDM optimiser, which outperforms ADAM on larger datasets (Pacal & Karaboga, 2021). Six networks underwent testing with three optimisers, each employing a learning rate of $1x10^{-3}$ and trained for 50 epochs. Subsequently, the same networks were tested with a learning rate of $1x10^{-3}$ over 100 epochs. The resulting outcomes are presented in Figures 5.1 and 5.2.

The preliminary experiments employed a learning rate of $1x10^{-3}$, and the outcomes are documented in Table 5.2. Later experiments used a bending ratio of $5x10^{-3}$ and 50-100 epochs. ADAM outperformed other optimisers with accuracy and recall of 0.87 and 0.86, respectively, across all six models. RMSprop is consistently the worst optimiser, with only 0.53 accuracy and 0.51,0.39 recall. ADAM is the best choice among the three options, with high sensitivity and better results. The results show that ADAM was the most efficient optimiser. Furthermore, 100 epochs produced better results with 0.87 accuracy and 0.86 recall compared to 50 epochs, which produced 0.80 accuracy and

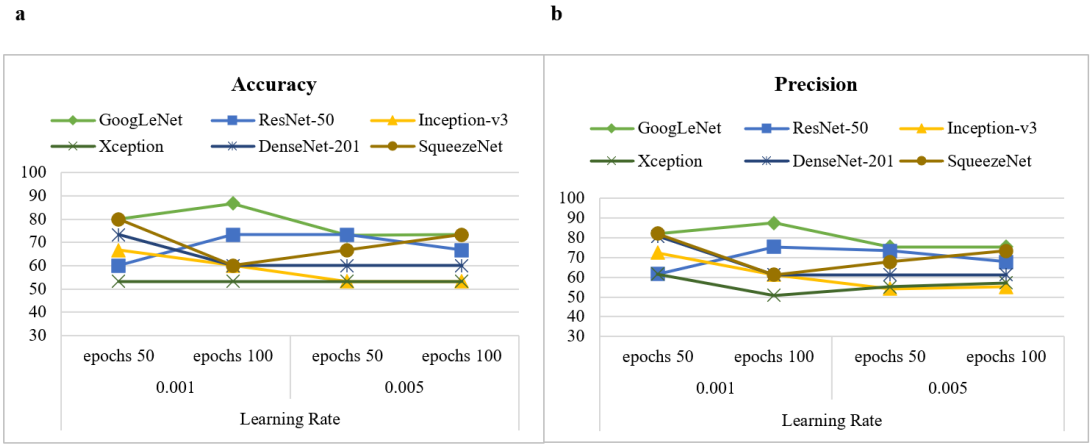

Figure 5.1: Accuracy and precision, optimiser: ADAM, LR: 0.001, epochs 50, 100

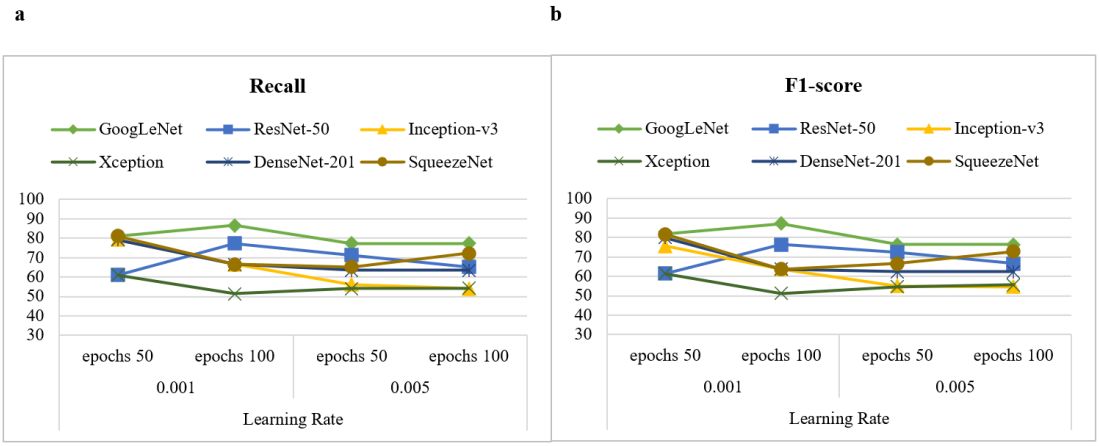

Figure 5.2: Recall and f1-score, optimiser: ADAM, LR: 0.001, 0.005, epochs: 50, 100

0.81 recall, and the $1x10^{-3}$ learning rate produced better results compared to $5x10^{-3}$, as shown in Table 5.3.

E 1.2: Performance on Balanced Data

The experiment aims to evaluate the proposed model on balanced data. Class imbalance can affect model performance. The imbalanced datasets used in this study deteriorated model performance. To address this issue, the imbalance was addressed by equalising

Table 5.2: Learning rate: 0.001 and 100 epochs

Model(s)	ADAM				SGDM				RMSprop			
	Acc.	**Prec**	**Recall**	**F1 score**	**Acc.**	**Prec**	**Recall**	**F1-score**	**Acc.**	**Prec**	**Recall**	**F1 score**
GoogLeNet	**86.7**	**87.6**	**86.6**	**87.1**	73.3	75.4	77.2	76.3	60.0	61.2	66.5	63.8
ResNet-50	73.3	75.4	77.2	76.3	53.3	50.8	51.4	51.1	53.3	50.8	51.4	51.1
Inception-v3	60.0	61.2	66.5	63.8	60.0	61.2	66.5	63.7	60.0	61.2	66.5	63.7
Xception	53.3	50.8	51.3	51.1	53.3	50.8	51.3	51.1	53.3	50.8	51.3	51.1
DenseNet-201	60.0	61.2	66.5	63.8	66.6	77.7	65.2	70.9	60.0	61.2	66.5	63.8
SqueezeNet	60.0	61.2	66.5	63.75	60.0	61.2	66.5	63.7	53.3	50.8	51.3	51.1

Table 5.3: Results with ADAM optimiser, 0.001 and 0.005 learning rate, and 100 epochs

Model(s)	0.001				0.005			
	Accuracy	**Precision**	**Recall**	**F1 score**	**Accuracy**	**Precision**	**Recall**	**F1 score**
GoogLeNet	**86.7**	**86.6**	**86.6**	**87.1**	75.4	75.4	77.2	76.3
ResNet-50	73.3	77.2	77.2	76.3	67.7	67.8	65.3	66.5
Inception-v3	60.0	66.5	66.5	63.7	55.1	55.2	54.1	54.6
Xception	53.3	51.3	51.4	51.1	57.2	57.2	54.2	55.6
DenseNet-201	60.0	66.5	66.5	63.8	61.2	61.1	63.5	62.3
SqueezeNet	60.0	66.1	66.5	63.5	73.3	73.4	72.2	72.8

the images across all classes, as demonstrated in Table 5.4. Following the oversampling of the data, we adjusted pre-trained networks and augmented a limited number of images for rotation and scaling before applying the DL algorithm. Similar parameters were employed to identify the most efficient deep learning architecture with the best hyperparameters. To find the best one, we tested three optimisers - ADAM, SGDM, and RMSprop - on both oversampled and augmented data. We compared the results by changing other parameters and selecting the best optimiser, which was then evaluated. The same optimiser and number of epochs were used to determine the optimal learning rate from $1x10^{-3}$ to $5x10^{-3}$. Finally, the most effective architecture was determined based

on the selected parameters. The findings are displayed in Figures 5.3 and 5.4.

Table 5.4: Number of polyps per category

	Adenomatous Lesions	**Hyperplastic Lesions**	**Serrated Lesions**
Benchmark Dataset	40	21	15
Balanced Dataset	40	36	33

The SGDM optimiser yielded the most favourable results for evaluation metrics, achieving an accuracy of 0.93 and a recall of 0.96 across various network configurations, surpassing other optimisers. Additional testing demonstrated that using SGDM with 100 epochs produced better results than with 50 epochs, irrespective of the learning rate, as shown in Table 5.5. Overall, the results confirmed that SGDM is the most efficient optimiser across all experimental settings.

Choosing the right learning rate is also important in training a DL model. In this experiment, an SGDM optimiser was used with 100 epochs, comparing results from two learning rates of $1x10^{-3}$ and $5x10^{-3}$, which showed that $1x10^{-3}$ gave an improved outcome. Class imbalance affects performance, but it can be handled before training for higher accuracy, as demonstrated in Table 5.6. Balancing data improves predictions. Results were generated for two and three-class classifications, as shown in Table 5.7. GoogLeNet achieved 0.87 accuracy, with 0.86 precision, 0.86 recall, and 0.87 F1-score on benchmark data using ADAM optimisers, a learning rate $1x10^{-3}$, and 100 epochs. ResNet-50 demonstrates high accuracy when applied to balanced datasets. SGDM is the better optimiser with 0.93 precision, 0.95 recall, and 0.94 F1 score.

A publication (Mesejo et al., 2016) used the benchmark dataset to classify polyps into two categories: hyperplasic and adenomatous. We experimented with these classes

using optimal hyperparameters, and the optimal outcomes are reported for comparison. Based on the findings presented in Table 5.7, ResNet-50 demonstrated the highest accuracy across benchmark and balanced datasets. These results answer the research question **RQ1b**. The ADAM and SGDM optimisers achieved comparable and top performance with a learning rate of 0.001 over 100 epochs, yielding an accuracy of 0.86, precision of 0.91, recall of 0.92, and F1 score of 0.91 that answers the research question **RQ1a**. However, the least effective CNN architectures for two-class classification are DenseNet201, achieving 0.40 accuracy, 0.66 precision, 0.37 recall, and 0.47 F1 score on benchmark data, and GoogLeNet, achieving 0.67 accuracy, 0.72 precision, 0.80 recall, and 0.76 F1 score on oversampled data.

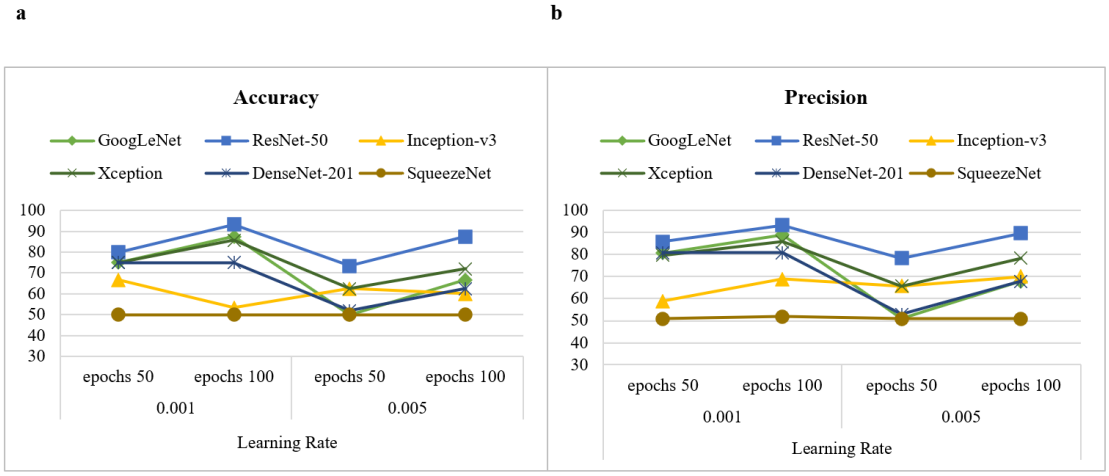

Figure 5.3: Accuracy and precision, optimisers: SGDM, LR: 0.001, 0.005, Epochs: 50, 100

5.1.2 Ensemble-Based Polyp Classification

This section examines the performance of a weighted average ensemble approach in classifying colorectal polyp images into adenoma, hyperplasia, and adenocarcinoma categories. Our goal is to enhance the classification system's generalisation capability

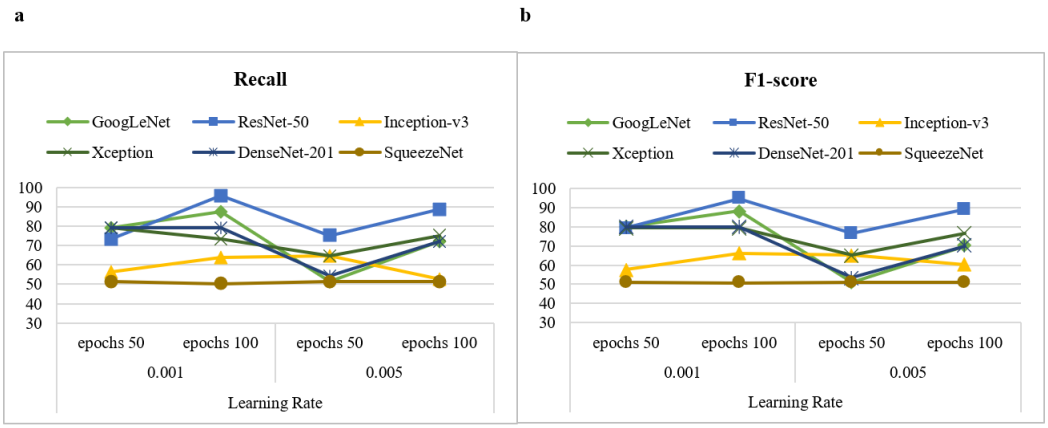

Figure 5.4: Recall and f1-score, optimiser: SGDM, LR: 0.001, 0.005, Epochs: 50, 100

Table 5.5: Results with 0.001 learning rate and 100 epochs

Model(s)	ADAM				SGDM				RMSprop			
	Acc.	**Prec**	**Recall**	**F1 score**	**Acc.**	**Prec**	**Recall**	**F1 score**	**Acc.**	**Prec**	**Recall**	**F1 score**
GoogLeNet	25.0	58.0	51.1	54.4	87.5	88.9	87.6	88.2	25	58.01	51.1	54.5
ResNet-50	53.3	48.6	48.6	48.6	**93.3**	**93.3**	**95.8**	**94.6**	60.0	60.0	63.8	61.8
Inception-v3	62.5	66.7	66.7	66.6	53.3	68.9	63.8	66.3	66.7	73.8	68.1	70.8
Xception	75.0	80.8	79.3	80.0	85.7	85.9	73.6	79.3	75.0	80.8	79.3	80.0
DenseNet-201	65.0	61.7	61.1	61.4	75.0	80.8	79.3	80.0	75.0	80.7	79.3	80.0
SqueezeNet	25.0	58.1	51.1	54.4	87.5	88.9	87.6	88.2	25.0	58.0	51.2	54.4

in handling the complex nature of polyps through ensemble learning. The ensemble method proposed leverages DL and involves two training stages. In the first stage, three independent pre-trained CNNs — GoogLeNet, ResNet-50, and Xception are individually trained using ImageNet as their base classifiers. In the subsequent training phase, an averaging-based ensemble learning approach is utilised to classify the input data. The weights derived from the base classifiers are averaged to determine the final decision.

Table 5.6: Results with SGDM optimiser and 100 epochs

Model(s)	0.001				0.005			
	Acc.	**Precision**	**Recall**	**F1-score**	**Acc.**	**Precision**	**Recall**	**F1-score**
GoogLeNet	87.5	88.9	87.5	88.2	66.7	67.9	72.2	70.0
ResNet-50	**93.3**	**93.3**	**95.8**	**94.6**	87.5	89.6	88.6	89.1
Inception-v3	53.3	68.9	63.8	66.3	60.0	70.0	52.8	60.2
Xception	85.7	85.9	73.6	79.3	72.0	78.3	75.2	76.7
DenseNet-201	75.0	80.8	79.8	80.0	62.5	67.9	72.2	70.0
SqueezeNet	50.0	51.8	50.4	51.1	50	50.8	51.4	51.1

Table 5.7: Optimized hyperparameter settings

No. of classes	Dataset	Model	optimiser	No of epochs	Learning Rate	Acc.	Precision	Recall	F1 score
2 classes	**Benchmark**	ResNet-50	ADAM	100	0.001	86.67	91.7	91.67	91.7
	Balanced	ResNet-50	SGDM	100	0.001	86.67	91.7	91.67	91.7
3 classes	**Benchmark**	GoogLeNet	ADAM	100	0.001	87.5	86.63	86.63	87.12
	Balanced	**ResNet-50**	**SGDM**	**100**	**0.001**	**93.3**	**93.33**	**95.83**	**94.56**

E 1.3: Ensemble Learning of Optimised Networks

Neural networks are notorious for their high variance and low bias, which can make them less reliable than desired. However, ensemble learning can assist in solving this problem. Ensemble learning introduces bias to the model by training multiple models and combining their predictions. This reduction in variance ultimately leads to more accurate predictions. Not only does this approach improve the reliability of the model, but it also enhances its performance. As a result, the model becomes more robust, producing predictions that are less influenced by the specifics of the training data and the training methodology. The results are depicted in Figure 5.5.

Ensemble learning involves the use of different models and training data to

Table 5.8: Performance evaluation of multiclass imbalanced and balanced dataset

Dataset	Model(s)	Data type	Accuracy	Precision	Recall	F1-score	Specificity	Sensitivity	TP	TN	FP
UCI Dataset	GoogLeNet	Imbalanced	86.7	87.6	86.6	87.1	0.87	0.84	65	141	10
	ResNet-50		73.3	75.4	77.2	76.3	0.71	0.78	56	132	24
	GoogLeNet	Balanced	87.5	88.9	87.2	87.6	0.87	0.87	95	204	14
	ResNet-50		92.7	91.8	93.5	92.7	0.92	0.93	101	210	9
PICCOLO Dataset	GoogLeNet	Imbalanced	73.1	75.2	72.4	73.2	0.75	0.71	255	593	84
	Xception		72.3	75.2	71.2	73.4	0.75	0.71	265	598	86
	ResNet-50		73.1	78.1	69.2	73.3	0.78	0.69	244	577	72
	GoogLeNet	Balanced	72.1	73.3	71.1	72.4	0.73	0.71	240	573	90
	Xception		72.4	71.1	73.2	72.3	0.71	0.73	239	572	98
	ResNet-50		73.2	73.4	72.3	73.3	0.73	0.72	242	575	90

generate an average of the predictions from those models. Assigning appropriate weights to each model's predictions can enhance the overall accuracy of this average. In the context of this study, this method is known as a weighted average set or model blending (Brownlee, 2018). To enhance the classification system's generalisation for complex colorectal polyps, we chose the top two optimised pre-trained networks—GoogLeNet and ResNet-50—based on a predefined accuracy threshold. Several high-performing deep learning networks are combined to enhance classification performance. This is achieved by training individual base classifiers with the ImageNet database. Combining the capabilities of these networks into an ensemble compensates for their limited capability to capture data distribution, resulting in a more accurate outcome. We improved the classifier's performance by developing an averaging ensemble-based classifier. We selected the optimal base classifier weight values through a grid search to maximise the ensemble model's performance.

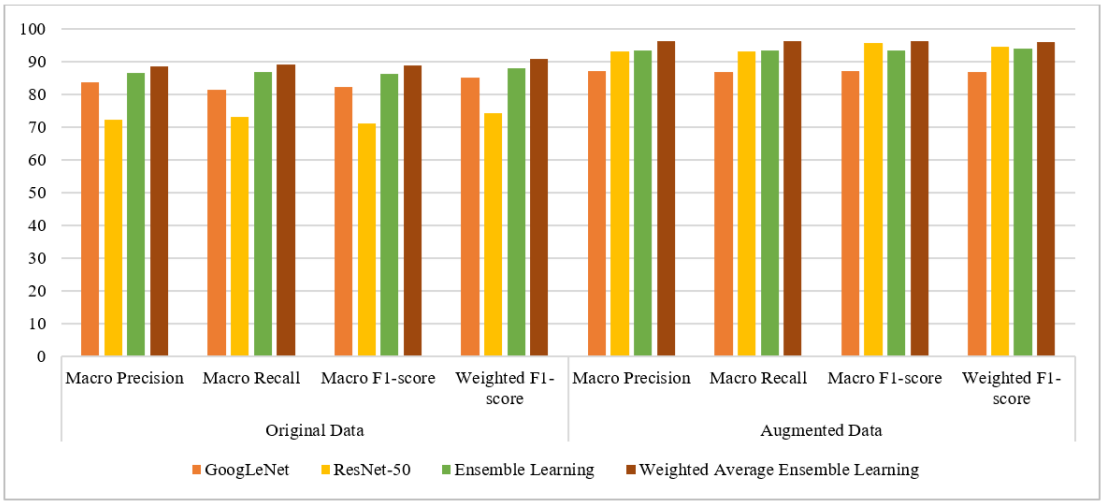

Figure 5.5: Weighted ensemble learning macro results comparison for imbalanced and balanced

E 1.4: Performance of the Base Classifier

During the initial phase of experiments, the original imbalanced dataset was utilized. Table 5.9 - 5.12 presents the results of the base classifiers on the validation and test sets. The F1-score measure was employed to assess the training potential of the base classifiers. Recall and precision were taken into account to assess the variety among the learners. Each of the base classifiers demonstrates learning proficiency, achieving F1 scores between 0.71 and 0.74 on the validation set and maintaining a consistent score of 0.73 on the test set. However, a thorough examination of the results shows that with the imbalanced dataset, the validation precision is notably low, reaching a value of 0.61

In this study, multiclass classification was conducted, considering the imbalance in the data. Performance metrics such as macro precision, macro recall, macro F1-score, and weighted F1-score were employed to assess performance across specific categories. The highest macro F1-score attained was 0.72 on the validation set with GoogLeNet and

0.75 on the test set with ResNet-50, as detailed in Table 5.13.

E 1.5: Performance of the Proposed Classifier

A deep ensemble learning classifier has been developed to address the intricate structure of colorectal polyps. Virtual biopsy is a sensitive and complex task that demands an accurate polyp classification. Hence, the ensemble learning method was devised to enhance polyp classification.

For the imbalanced dataset, the macro F1 scores achieved by the average and weighted-average ensemble models were 0.70 and 0.71 on the validation set and 0.73 and 0.74 on the test set, respectively, using the PICCOLO dataset. These findings indicate that ensemble-averaging learning did not enhance the outcomes compared to the base classifiers. Nevertheless, the quantitative assessment of the average and weighted average ensemble classifiers indicates that allocating an appropriate blend of weights produces encouraging outcomes and outperforms individual base learners and average ensemble architecture. Furthermore, data augmentation has indicated improved performance, achieving 0.76 and 0.79 in the validation and 0.76 and 0.84 on the test set using the PICCOLO dataset. The findings are displayed in Table 5.14. The macro and weighted F1 scores demonstrate that the base classifiers successfully captured the complex representations of various types of polyps. A similar trend is observed for the CPDC dataset, where the proposed approach achieved an accuracy of 0.75 on the test and 0.76 on the validation set.

The effectiveness of several pre-trained CNNs with diverse model designs was assessed to handle the challenge of CRC classification. However, these classifiers failed to achieve exceptional results on colonoscopy images, unlike the proposed technique. On

the other hand, leveraging the collective strength of weak learners has led to significant improvements in results. Furthermore, assigning appropriate weights to the base learners has notably enhanced image classification. Figure 5.6 illustrates the comparison of F1 scores.

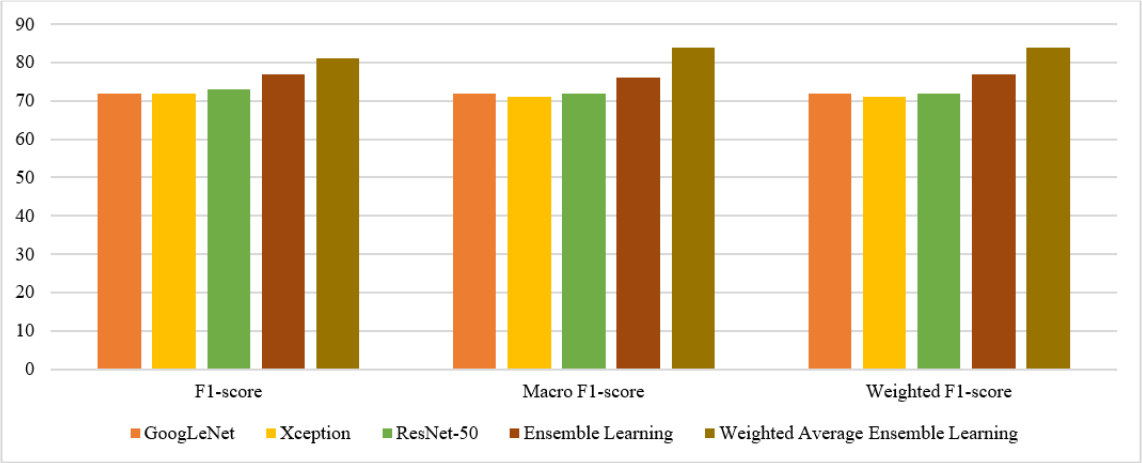

Figure 5.6: Base and ensemble classifier performance on balanced dataset

The proposed method demonstrates a 2% increase in macro F1-score on the validation set and 3% on the test set for original data. In addition, a notable improvement of 4% on the validation set and 12% on the test set is observed compared to the highest value achieved by individual base classifiers. Furthermore, the proposed weighted ensemble learning notably improves macro precision in the validation and test sets; Figure 5.14 illustrates the results produced.

For the imbalanced dataset, the weighted-average ensemble models achieved average macro-F1 scores of 0.73 and 0.74 on the test set. These findings indicate that average ensemble learning does not offer improvement compared to using individual base classifiers. On the other hand, assigning suitable weights to base classifiers yields better performance than using single or average ensemble models. Furthermore, the

balanced data exhibited superior performance, achieving scores of 0.76 in the validation set and 0.79 on the test set.

We evaluated multiple pre-trained CNNs with different architectures for classifying colorectal polyps, but they did not work well on colonoscopy images. However, combining weak learners improved the results, and assigning appropriate weights to the base classifiers significantly improved the results. Our approach demonstrated a 3% improvement in the macro F1 score in the test set compared to the benchmark data and a notable 12% increase over the highest score achieved by each individual base classifier.

The results of the validation and test set are presented in Tables 5.14 - 5.17. The ensemble learning approach demonstrates robust capabilities in polyp classification. The performance of the base, average ensemble, and weighted average ensemble is validated on the validation set and tested on the test set.

Table 5.9: Performance of the base and ensemble classifier on imbalanced dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score
GoogLeNet		0.72	0.74	0.73	0.73
Xception		0.71	0.75	0.72	0.72
ResNet-50	Test	0.73	0.75	0.69	0.73
Ensemble learning		0.73	0.78	0.73	0.76
Ensemble learning (Weighted Average)		**0.74**	**0.78**	**0.73**	**0.75**
GoogLeNet		0.73	0.68	0.77	0.75
Xception		0.70	0.61	0.79	0.70
ResNet-50	Validation	0.72	0.63	0.79	0.72
Ensemble learning		0.74	0.64	0.83	0.73
Ensemble learning (Weighted Average)		**0.76**	**0.69**	**0.84**	**0.75**

Table 5.10: Sensitivity and specificity of the base and ensemble classifier on imbalanced dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet		0.73	0.70	255	593	84	107
Xception		0.70	0.69	265	598	86	110
ResNet-50	Test	0.74	0.71	244	577	72	106
Ensemble Learning		0.75	0.72	235	558	68	108
Ensemble learning (Weighted Average)		**0.76**	**0.73**	**244**	**572**	**70**	**98**
GoogLeNet		0.68	0.81	662	1559	314	156
Xception		0.65	0.73	622	1519	398	152
ResNet-50	Validation	0.68	0.75	642	1540	358	154
Ensemble learning		0.69	0.76	665	1562	372	92
Ensemble learning (Weighted Average)		**0.70**	**0.78**	**695**	**1622**	**367**	**97**

Table 5.11: Performance of the base and ensemble classifier on augmented dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score
GoogLeNet		0.73	0.71	0.72	0.72
Xception		0.71	0.70	0.73	0.70
ResNet-50	Test	0.72	0.72	0.72	0.73
Ensemble learning		0.76	0.77	0.74	0.73
Ensemble learning (Weighted Average)		**0.80**	**0.81**	**0.82**	**0.80**
GoogLeNet		0.69	0.69	0.68	0.68
Xception		0.73	0.74	0.73	0.73
ResNet-50	Validation	0.74	0.76	0.73	0.75
Ensemble learning		0.77	0.76	0.76	0.77
Ensemble learning (Weighted Average)		**0.81**	**0.82**	**0.80**	**0.81**

E.1.6: Comparative Analysis of Classification Performance

The limited use of this publicly available dataset makes it difficult to compare with other methods. Two studies have previously classified colorectal polyps on this benchmark dataset. Zhang et al. (2016) proposed a transfer learning framework based on CNN, where features learned from a non-medical dataset were utilised. It is crucial to evaluate

Table 5.12: Sensitivity and specificity of the base and ensemble classifier on augmented dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet		0.72	0.70	240	573	90	96
Xception		0.71	0.72	239	572	98	90
ResNet-50	Test	0.72	0.73	242	575	90	92
Ensemble learning		0.78	0.75	253	591	66	89
Ensemble learning (Weighted average)		**0.82**	**0.81**	**270**	**603**	**61**	**65**
GoogLeNet		0.70	0.69	975	2346	396	396
Xception		0.73	0.73	1018	2389	353	353
ResNet-50	Validation	0.74	0.74	1039	2410	332	332
Ensemble Learning		0.76	0.75	1056	2427	315	315
Ensemble learning (Weighted average)		**0.81**	**0.80**	**1122**	**2493**	**249**	**249**

Table 5.13: Performance evaluation of multiclass augmented dataset

Models	Dataset	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score
GoogLeNet		0.71	0.70	0.72	0.71
Xception		0.70	0.69	0.70	0.71
ResNet-50	Test	0.71	0.72	0.70	0.70
Ensemble Learning		0.74	0.74	0.73	0.74
Ensemble Learning (Weighted Average)		**0.81**	**0.81**	**0.84**	**0.84**
GoogLeNet		0.69	0.70	0.69	0.71
Xception		0.70	0.71	0.72	0.73
ResNet-50	Validation	0.72	0.73	0.75	0.76
Ensemble learning		0.75	0.74	0.76	0.75
Ensemble learning (Weighted Average)		**0.80**	**0.79**	**0.79**	**0.81**

our results in the context of binary classification in order to compare them with the results of their investigation. As a result of our experiments, we were able to classify colorectal

Table 5.14: Performance of the base and ensemble classifier on binary class augmented dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score	MCC
GoogLeNet		0.72	0.73	0.72	0.73	0.44
Xception		0.70	0.72	0.74	0.73	0.36
ResNet-50	Test	0.71	0.73	0.75	0.74	0.43
Ensemble Learning		0.73	0.73	0.73	0.73	0.45
Ensemble Learning (Weighted Average)		**0.75**	**0.76**	**0.76**	**0.76**	**0.47**
GoogLeNet		0.71	0.72	0.71	0.71	0.44
Xception		0.69	0.68	0.70	0.69	0.34
ResNet-50	Validation	0.72	0.71	0.73	0.72	0.44
Ensemble learning		0.74	0.74	0.74	0.74	0.46
Ensemble learning (Weighted Average)		**0.76**	**0.75**	**0.77**	**0.76**	**0.49**

polyps as either hyperplastic or adenomatous using an optimised hyperparameter configuration. The work achieved 85.9% accuracy, 0.87 precision, 0.87 recall, and 0.87.0 F1 scores in a two-class classification task. Our proposed approach of creating a weighted average ensemble improved the classifier's performance by 2% on imbalanced data and 3% on balanced data. Our framework outperformed their approach by achieving 86.6% accuracy, 91.7% precision, recall, and F1-score with both benchmark and oversampled data.

The other comparative evaluation was accomplished by Mesejo et al. (2016), where in order to implement virtual biopsy and classify colorectal polyps into hyperplastic lesions, serrated adenomas, and adenomas, a three-class classification framework was developed by combining machine learning and computer vision algorithms. The research incorporated Random Forest (RF), Random Subspace (RS), and Support Vector Machine (SVM) classifiers to achieve its goal. The approach yielded an accuracy of 82.46%,

sensitivity of 72.74%, and specificity of 85.88%. After comparing the outcomes of our three-class classification, it became evident that our proposed framework surpasses traditional machine learning and deep learning approaches. The framework produced an accuracy of 90.1% compared to 0.96 and a recall of 0.92 compared to 0.97 on benchmark data and oversampled data, respectively.

This research explores the effectiveness of various deep learning models for colorectal polyp classification, including GoogleNet, ResNet-50, ensemble learning, and weighted average ensemble learning. Afterwards, the results are compared to published studies regarding polyp classification using deep learning models. The CNN model proposed by Hsu et al. (2021) achieved the highest accuracy of 82.8%, as observed. Further, Jae et al. (2021) proposed a transfer learning model using AlexNet as the backbone. In terms of accuracy, the model performed best with fully connected networks, achieving a high score of 0.79. However, this study surpassed previous investigations by achieving even higher accuracy rates of 0.96 and 0.91 on balanced and imbalanced datasets, respectively, through approach of weighted averaging.

5.1.3 Reliability Analysis

Figure 5.9 shows the comparison of error rates and kappa coefficients for imbalanced and balanced data sets of the UCI dataset. The kappa values for the base classifiers, GoogLeNet and ResNet-50, are 0.79 and 0.58 on the benchmark data, respectively. The kappa values improve to 0.81 and 0.89 with balanced data. However, in terms of ensemble classifiers, the average ensemble generates a 0.90 kappa value, and the weighted ensemble further improves the result to 0.94.

Figure 5.10 compares the kappa coefficient and the error values performed on the

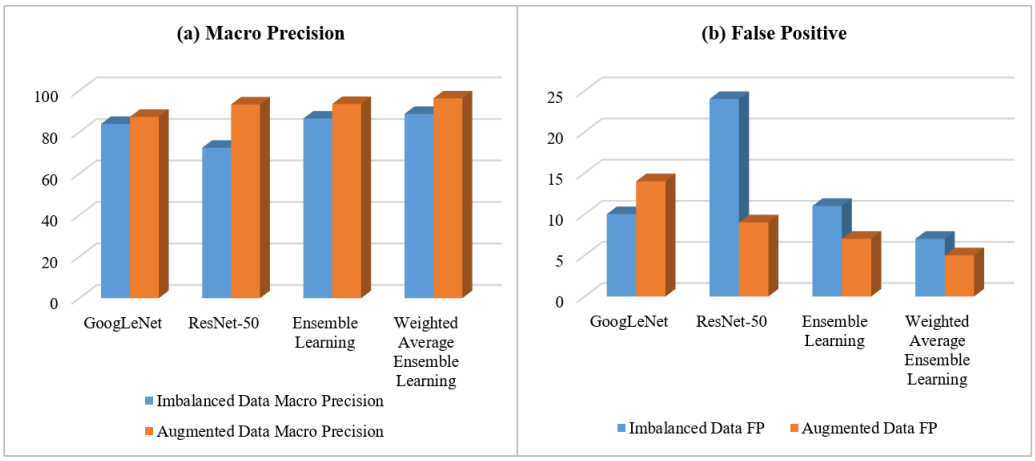

Figure 5.7: Macro precision and false positive rate comparison of imbalanced and balanced UCI dataset

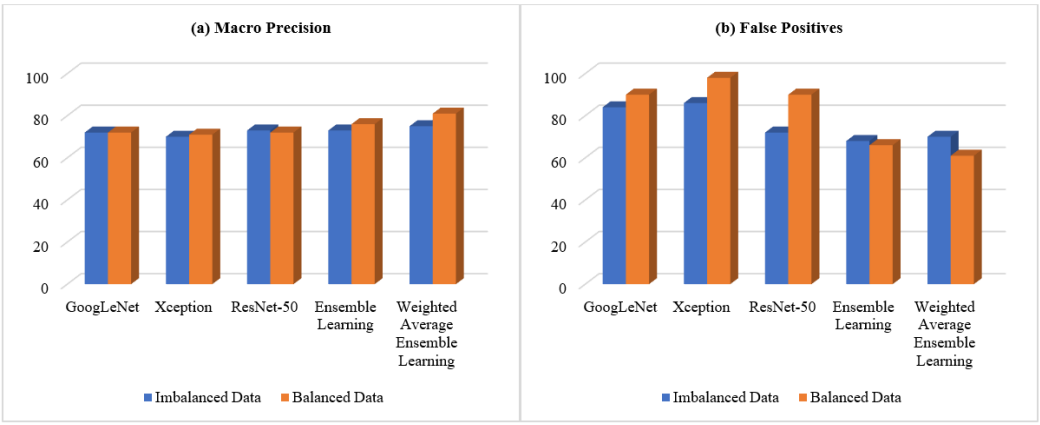

Figure 5.8: Macro precision and false positive rate comparison of imbalanced and balanced PICCOLO dataset

PICCOLO dataset. The base classifiers GoogLeNet, Xception, and ResNet-50 had kappa values of 0.59, 0.55, and 0.59, respectively. In terms of ensemble classifiers, the average ensemble produces a kappa value of 0.61. However, the weighted ensemble improves the result to 0.62, illustrating that the model error decreases in both scenarios as the kappa coefficient increases. This proposed ensemble method demonstrates an acceptable level of reliability, as evidenced by a significant increase in the kappa coefficient. Figure 5.11

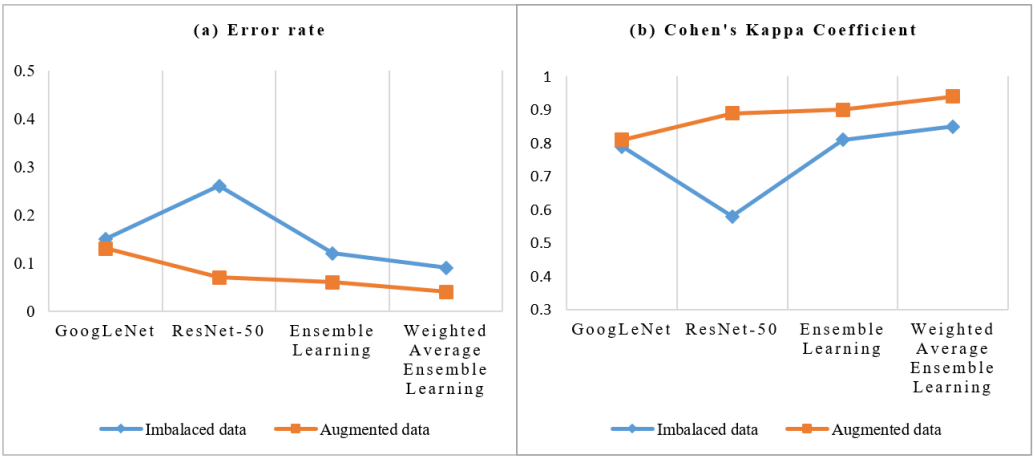

Figure 5.9: Reliability comparison using Cohen's kappa coefficient on UCI dataset

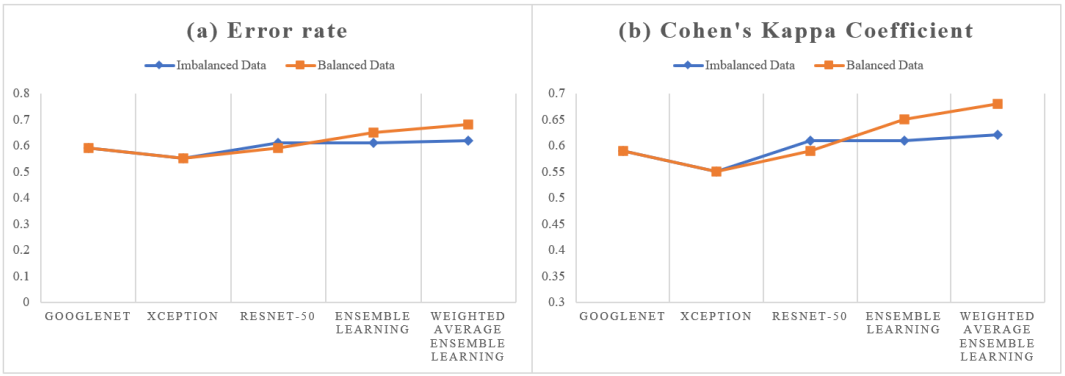

Figure 5.10: Reliability comparison using Cohen's kappa coefficient on PICCOLO dataset

shows the ROC-AUC value of 0.94, indicating good model separability for the PICCOLO dataset, and Figure 5.12 shows the ROC-AUC as 0.89 for the two-class classification and 0.91 for the three-class classification in the UCI dataset. These values indicate that the proposed model has a high degree of separability.

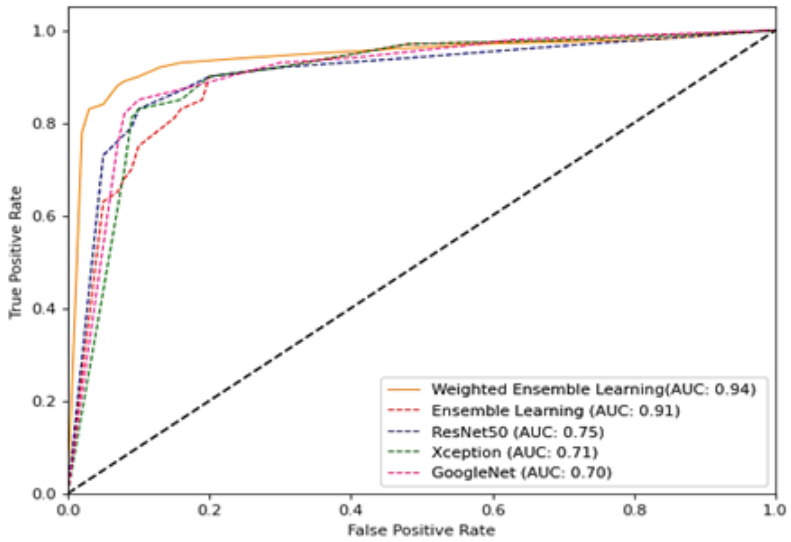

Figure 5.11: AUC-ROC of PICCOLO Dataset

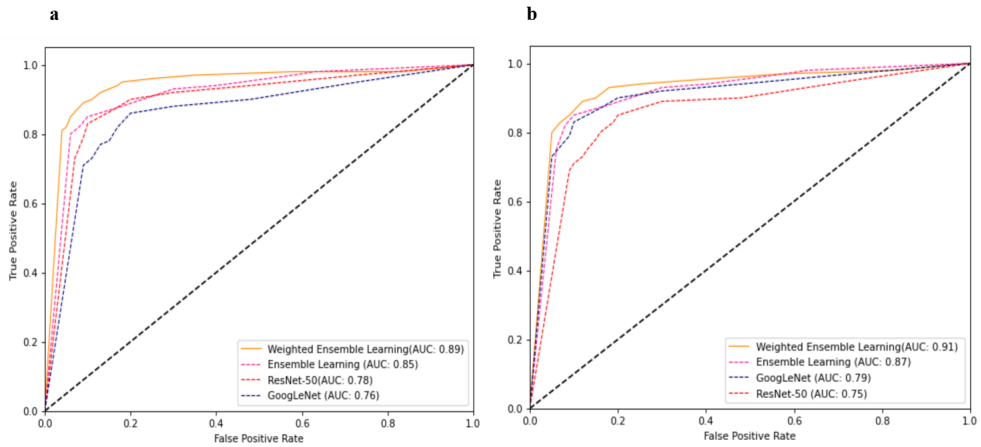

Figure 5.12: AUC-ROC of binary and multiclass classification on UCI dataset

5.1.4 Precision-Recall Analysis

Besides evaluating the model's sensitivity, it is crucial to assess the precision of the proposed system. Precision measures the proportion of correctly identified positive cases among all cases predicted as positive in the dataset. A few false positives can significantly impact the precision of a CAD system, especially in the context of imbalanced data, resulting in a reduced F1 score. In medical applications, data often exhibit class imbalance,

where one class predominates with more instances than another smaller but clinically significant class. Such systems generate a high false negative rate, misclassifying minority instances as the majority class (Yang et al., 2011). In such systems, misclassifying positive class examples can cost highly and adversely affect system performance and diagnosis. The proposed system was designed to handle class imbalance and improve accuracy by reducing false positive and negative predictions. The system achieved a precision of 95.5 in the UCI dataset for balanced data, indicating its ability to detect positive cases, as shown in Figure 5.7. The comparison of the precision and false positive macros of the proposed approach is presented based on both the imbalanced and balanced PICCOLO dataset, as shown in Figure 5.8. The system achieved a precision of 0.81 on this dataset for balanced data, indicating its ability to identify positive cases accurately.

Figure 5.13 compares the performance of base and ensemble classifiers on the augmented dataset, and Figure 5.14 compares macro precision in the proposed approach for both imbalanced and augmented datasets. Besides model sensitivity, analyzing precision is crucial for assessing the proposed system. Precision measures the correctly identified positive cases out of all positive instances in the data. The presence of a small fraction of false positive values in Figure 5.15 can significantly impact the precision of the CAD system, especially if the data is imbalanced, leading to a decrease in the F1-score. Figure. 5.16 presents the performance comparison of the CPDC dataset, whereas Figure 5.17 shows the False positive and Precision comparison of CPDC. In the medical domain, where data is usually imbalanced, this misclassification can affect the system's classification and have an adverse effect on diagnosis. The precision of the proposed system is 0.81 for augmented data, which indicates that the system can identify positive cases.

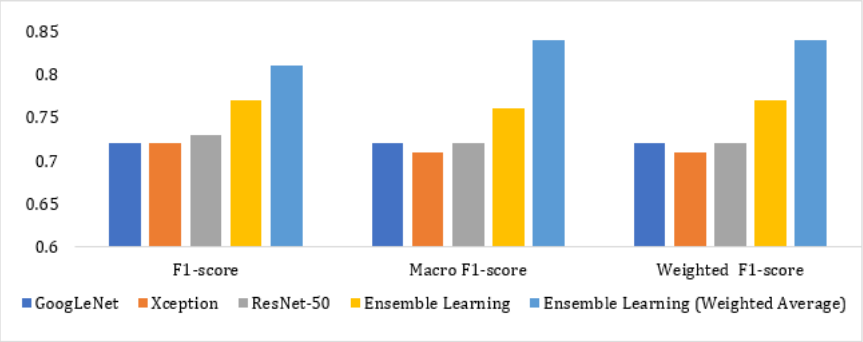

Figure 5.13: Base and ensemble classifier performance on augmented dataset

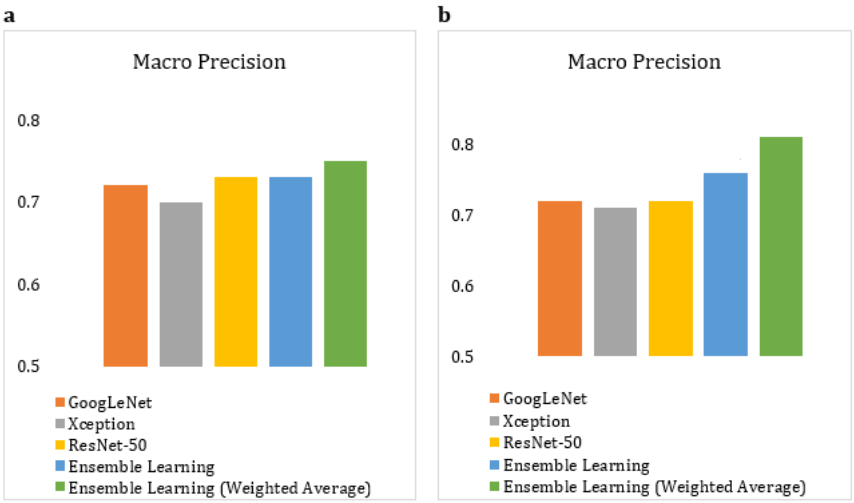

Figure 5.14: Precision-based comparison of imbalanced and augmented data

Figure 5.18 shows a comparison of validation and test set with Kappa coefficient and error values. The kappa values for the base classifiers GoogLeNet, Xception, and ResNet-50 are 0.59, 0.55, and 0.59, respectively. However, the average ensemble achieves a kappa value of 0.61, and the weighted ensemble further improves the result to 0.62. The graph demonstrates that as the Kappa coefficient increases, the model's error decreases in both scenarios. This notable increase in the Kappa coefficient indicates that the proposed ensemble method exhibits a satisfactory level of reliability. Figure 5.19 illustrates the ROC-AUC value of 0.94, indicating that the proposed model exhibits high

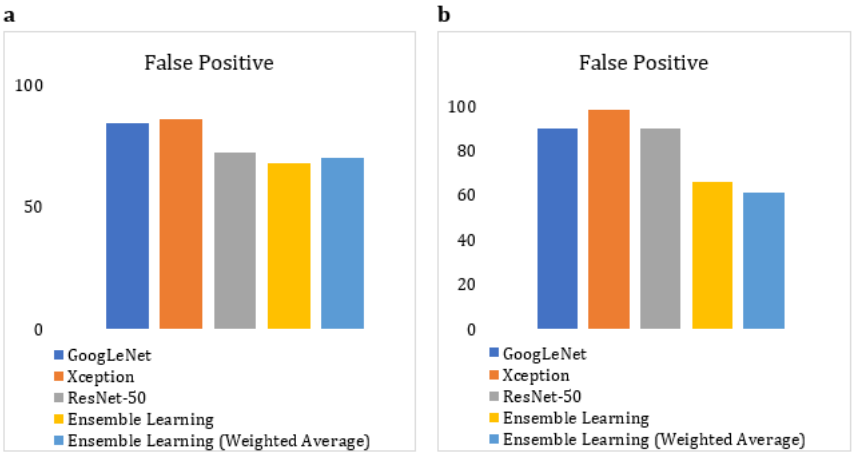

Figure 5.15: False positive rate comparison of imbalanced and augmented data

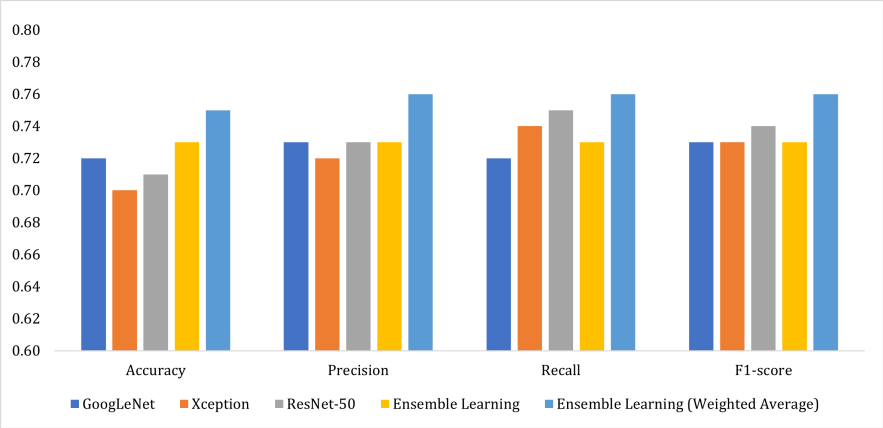

Figure 5.16: CPDC dataset performance analysis

separability. The Matthews Correlation Coefficient (MCC) is a fair evaluation metric for the imbalanced binary (CPDC) dataset. The results depicted in Figure 5.20 demonstrate an upward trend in the MCC value, indicating improved performance of the proposed models compared to the base classifiers. Table 5.18 presents the execution time and error of each architecture in both datasets used in this study.

The experiments in this study employed various deep learning models for the classification of colorectal polyps, including GoogleNet, ResNet-50, ensemble learning,

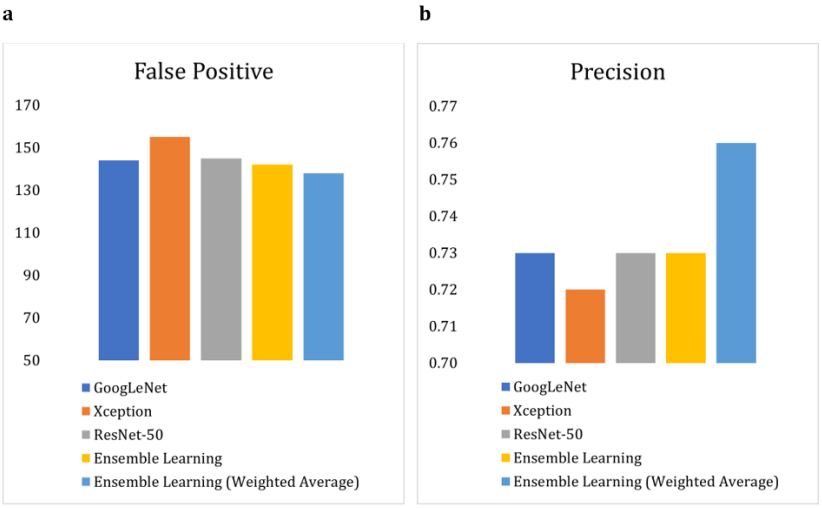

Figure 5.17: False positive and precision comparison of CPDC augmented data

Table 5.15: Performance evaluation of multiclass augmented dataset

Models	**Dataset**	**Macro Precision**	**Macro Recall**	**Macro F1-score**	**Weighted F1-score**
GoogLeNet	Test	0.71	0.70	0.72	0.71
Xception		0.70	0.69	0.70	0.71
ResNet-50		0.71	0.72	0.70	0.70
Ensemble Learning		0.74	0.74	0.73	0.74
Ensemble Learning (Weighted Average)		**0.81**	**0.81**	**0.84**	**0.84**
GoogLeNet	Validation	0.69	0.70	0.69	0.71
Xception		0.70	0.71	0.72	0.73
ResNet-50		0.72	0.73	0.75	0.76
Ensemble learning		0.75	0.74	0.76	0.75
Ensemble learning (Weighted Average)		**0.80**	**0.79**	**0.79**	**0.81**

and weighted average ensemble learning. The results are then compared with published works on polyp classification using deep learning models. The highest accuracy achieved in this study is 82.8% using the CNN model proposed by Hsu et al. (2021). Furthermore, Jae et al. (2021) utilised AlexNet as a backbone in transfer learning, achieving a maximum

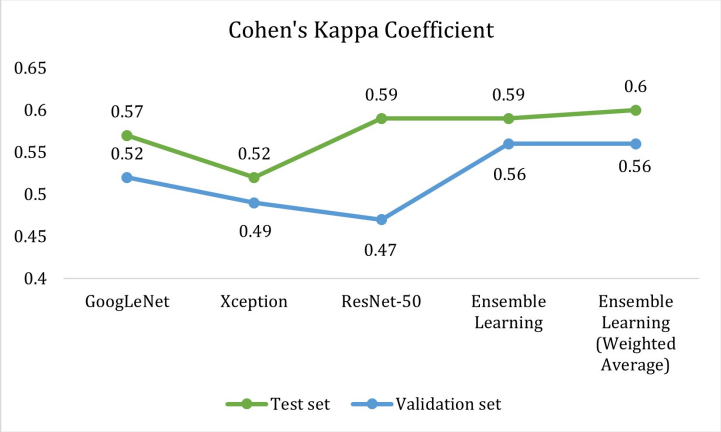

Figure 5.18: Reliability comparison using Cohen's kappa coefficient

Table 5.16: Performance of the base and ensemble classifier on binary class augmented dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score	MCC
GoogLeNet		0.72	0.73	0.72	0.73	0.44
Xception		0.70	0.72	0.74	0.73	0.36
ResNet-50	Test	0.71	0.73	0.75	0.74	0.43
Ensemble Learning		0.73	0.73	0.73	0.73	0.45
Ensemble Learning (Weighted Average)		**0.75**	**0.76**	**0.76**	**0.76**	**0.47**
GoogLeNet		0.71	0.72	0.71	0.71	0.44
Xception		0.69	0.68	0.70	0.69	0.34
ResNet-50	Validation	0.72	0.71	0.73	0.72	0.44
Ensemble learning		0.74	0.74	0.74	0.74	0.46
Ensemble learning (Weighted Average)		**0.76**	**0.75**	**0.77**	**0.76**	**0.49**

Table 5.17: Sensitivity and specificity of the base and ensemble classifier on binary class augmented dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet		0.59	0.56	558	256	144	139
Xception		0.43	0.50	540	245	155	180
ResNet-50	Test	0.43	0.50	545	245	145	140
Ensemble Learning		0.58	0.62	548	258	142	134
Ensemble Learning (Weighted Average)		**0.62**	**0.63**	**564**	**263**	**138**	**132**
GoogLeNet		0.59	0.59	538	249	141	138
Xception		0.64	0.65	526	210	180	140
ResNet-50	Validation	0.64	0.65	547	258	142	141
Ensemble learning		0.65	0.65	557	261	140	132
Ensemble learning (Weighted Average)		**0.67**	**0.69**	**558**	**302**	**138**	**127**

Table 5.18: Execution time comparison of datasets

Model(s)	Dataset	PICCOLO Dataset		CPDC Dataset	
		Error	Execution Time	Error	Execution Time
GoogLeNet		0.28	16min 55s	0.28	09min 25s
Xception		0.28	15min 04s	0.30	10min 14s
ResNet-50	Test	0.27	16min 03s	0.29	12min 22s
Ensemble Learning		0.23	21min 55s	0.27	14min 53s
Ensemble Learning (Weighted Average)		0.19	22min 34s	0.25	15min 04s
GoogLeNet		0.29	22min 11s	0.29	10min 13s
Xception		0.26	19min 31s	0.31	13min 33s
ResNet-50	Validation	0.24	18min 11s	0.28	13min 10s
Ensemble Learning		0.23	23min 04s	0.26	14min 02s
Ensemble Learning (Weighted Average)		0.18	24min 12s	0.24	15min 18s

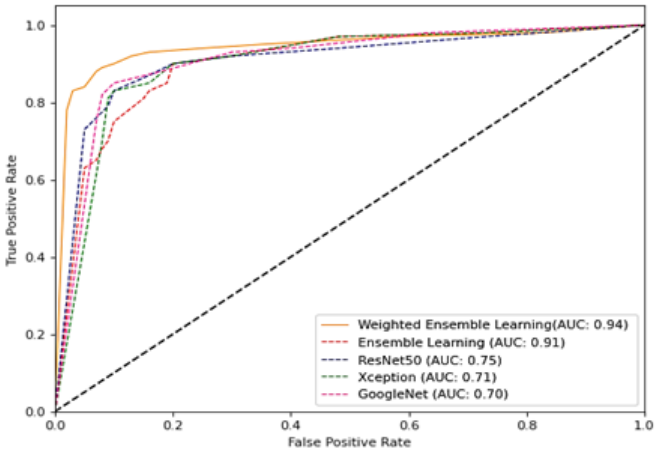

Figure 5.19: AUC-ROC of the proposed classifier on test data

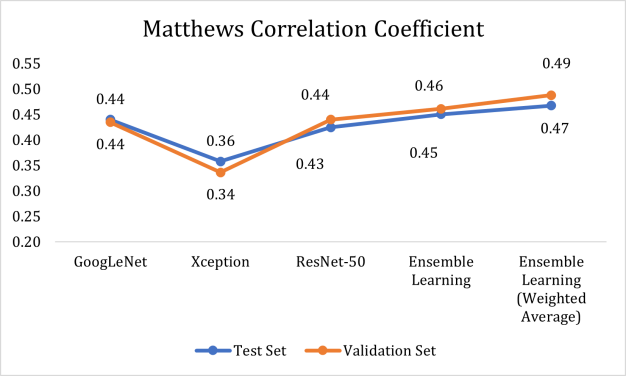

Figure 5.20: MCC comparison of validation and test set

accuracy of 79% with variations in fully connected networks. However, our method surpasses these recent works by achieving the highest accuracy of 96.3% and 90.5% on both balanced and imbalanced datasets using Weighted Average Ensemble Learning. A comparison with existing recent studies is presented in Table 5.19 and Table 5.20.

Table 5.19: Comparative analysis of UCI and PICCOLO datasets with existing studies

Dataset	No of Classes	Method		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
UCI Dataset	2	Transfer Learning (Mesejo et al., 2016)		85.9	87.3	87.6	86.0
		Imbalanced Dataset	Optimized Network	86.6	91.7	91.6	91.7
			Ensemble Learning	87.9	91.9	91.7	91.3
			Weighted Ensemble Learning	**88.1**	**92.4**	**91.9**	**91.5**
		Balanced Dataset	Optimized Network	86.7	91.7	91.6	91.7
			Ensemble Learning	88.3	91.9	92	91.8
			Weighted Ensemble Learning	**89.8**	**92.3**	**92.1**	**92.2**
	3	**SVM** (Zhang et al., 2016)		82.5	N.A.	72.7	N.A.
		Imbalanced Dataset	Optimized Network	86.7	86.6	86.6	87.1
			Ensemble Learning	88.2	88.9	88.3	88.6
			Weighted Ensemble Learning	**90.1**	**91.2**	**91.5**	**91.3**
		Balanced Dataset	Optimized Network	92.7	91.8	93.5	92.7
			Ensemble Learning	93.6	93.6	93.6	93.6
			Weighted Ensemble Learning	**96.3**	**95.5**	**97.2**	**96.3**
PICCOLO Dataset	3	Imbalanced Data	Optimized Network	73.1	78.1	69.2	73.3
			Ensemble Learning	73.2	78.4	69.1	73.4
			Weighted Ensemble Learning	**74.4**	**78.2**	**71.3**	**74.2**
		Balanced Dataset	Optimized Network	73.2	73.4	72.3	73.3
			Ensemble Learning	77.3	79.1	74.2	77.1
			Weighted Ensemble Learning	**81.2**	**82.4**	**81.1**	**81.3**
CVC-Clinic, CGMH-WL, NBI (Hsu et al., 2021)	3	N.A	CNN model-NBI	82.8	82.5	95.2	81.5
			CNN model-WL	72.2	75.3	88.7	81.5
Children's medicine department of Gachon University Gil Hospital in South Korea (Jae et al., 2021)	3	N.A	AlexNet+SOS,No transfer of fc6, fc7	0.706 ± 0.041	0.685 ± 0.077	0.807 ± 0.140	0.729 ± 0.052
			AlexNet+SOS,Transfer of fc6	0.761 ± 0.062	0.770 ± 0.101	0.793 ± 0.139	0.766 ± 0.061
			AlexNet+SOS,Transfer of fc6 and fc7	0.782 ± 0.037	0.737 ±0.061	0.893 ± 0.052	0.804 ± 0.027

Table 5.20: Comparative analysis of PICCOLO and CPDC datasets with existing studies

Dataset	Model(s)		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PICCOLO Dataset	GoogLeNet	Test	0.73	0.71	0.72	0.72
	Xception		0.71	0.70	0.73	0.70
	ResNet-50		0.72	0.72	0.72	0.73
	Ensemble Learning		0.76	0.77	0.74	0.73
	Ensemble Learning (Weighted Average)		**0.80**	**0.81**	**0.82**	**0.80**
	GoogLeNet	Validation	0.69	0.69	0.68	0.68
	Xception		0.73	0.74	0.73	0.73
	ResNet-50		0.74	0.76	0.73	0.75
	Ensemble Learning		0.77	0.76	0.76	0.77
	Ensemble Learning (Weighted Average)		**0.81**	**0.82**	**0.80**	**0.81**
CPDC Dataset	GoogLeNet	Test	0.72	0.73	0.72	0.73
	Xception		0.70	0.72	0.74	0.73
	ResNet-50		0.71	0.73	0.75	0.74
	Ensemble Learning		0.73	0.73	0.73	0.73
	Ensemble Learning (Weighted Average)		**0.75**	**0.76**	**0.76**	**0.76**
	GoogLeNet	Validation	0.71	0.72	0.71	0.71
	Xception		0.69	0.68	0.70	0.69
	ResNet-50		0.72	0.71	0.73	0.72
	Ensemble Learning		0.74	0.74	0.74	0.74
	Ensemble Learning (Weighted Average)		**0.76**	**0.75**	**0.77**	**0.76**
Children's medicine department of Gachon University Gil Hospital in South Korea (Jae et al., 2021)	AlexNet+SOS, No transfer of fc6, fc7		0.706 ± 0.041	0.685 ± 0.077	0.807 ± 0.140	0.729 ± 0.052
	AlexNet+SOS,Transfer of fc6		0.761 ± 0.062	0.770 ± 0.101	0.793 ± 0.139	0.766 ± 0.061
	AlexNet+SOS,Transfer of fc6 and fc7		0.782 ± 0.037	0.737 ± 0.061	0.893 ± 0.052	0.804 ± 0.027
	AlexNet+SOS,Addition of fc9		0.795 ± 0.045	0.766 ± 0.066	0.868 ± 0.069	0.809 ± 0.034
CVC-Clinic, CGMH-WL, NBI (Hsu et al., 2021)	CNN model-NBI		0.82	0.82	0.95	0.81
	CNN model-WL		0.72	0.75	0.88	0.81

Table 5.21: Evaluation of concordant cDCGAN on UCI, PICCOLO, and CPDC datasets

Method	Dataset	FID	IS	PSNR	SSIM
Concordant cGAN	UCI	4.5	7.8	20 dB	0.6
	PICCOLO	4.9	8.2	22 dB	0.7
	CPDC	5.6	7.5	23 dB	0.7

5.2 Synthetic Image Generation and Deep Feature Extraction

This section presents the results generated by SIGB and DFEB from the datasets. The imbalance between the classes significantly hinders the performance of the model. To address this problem, the imbalance was removed from the data by making the number of images in all classes equal through data augmentation. However, the images generated through cGAN do not always have the best quality and diversity for training an efficient model. Therefore, we performed two sets of experiments to determine the enhancement in classifier performance by incorporating a neural network that filters the concordant synthetic images for data augmentation. The first experiment includes data augmentation without incorporating the proposed concordant classifier and predicted class labels. The second experiment was conducted using the proposed concordant classifier, which generated and classified high-quality and diverse CRC data. The experiments on the datasets examined the performance of the concordant synthetic image dataset unified with feature extraction through ViT. The network architecture of the cDCGAN generator, the discriminator and their combined structure are shown in Figures 5.21, 5.22, and 5.23, respectively. The subsections 5.2.1 to 5.2.4 discuss the experimental results in detail.

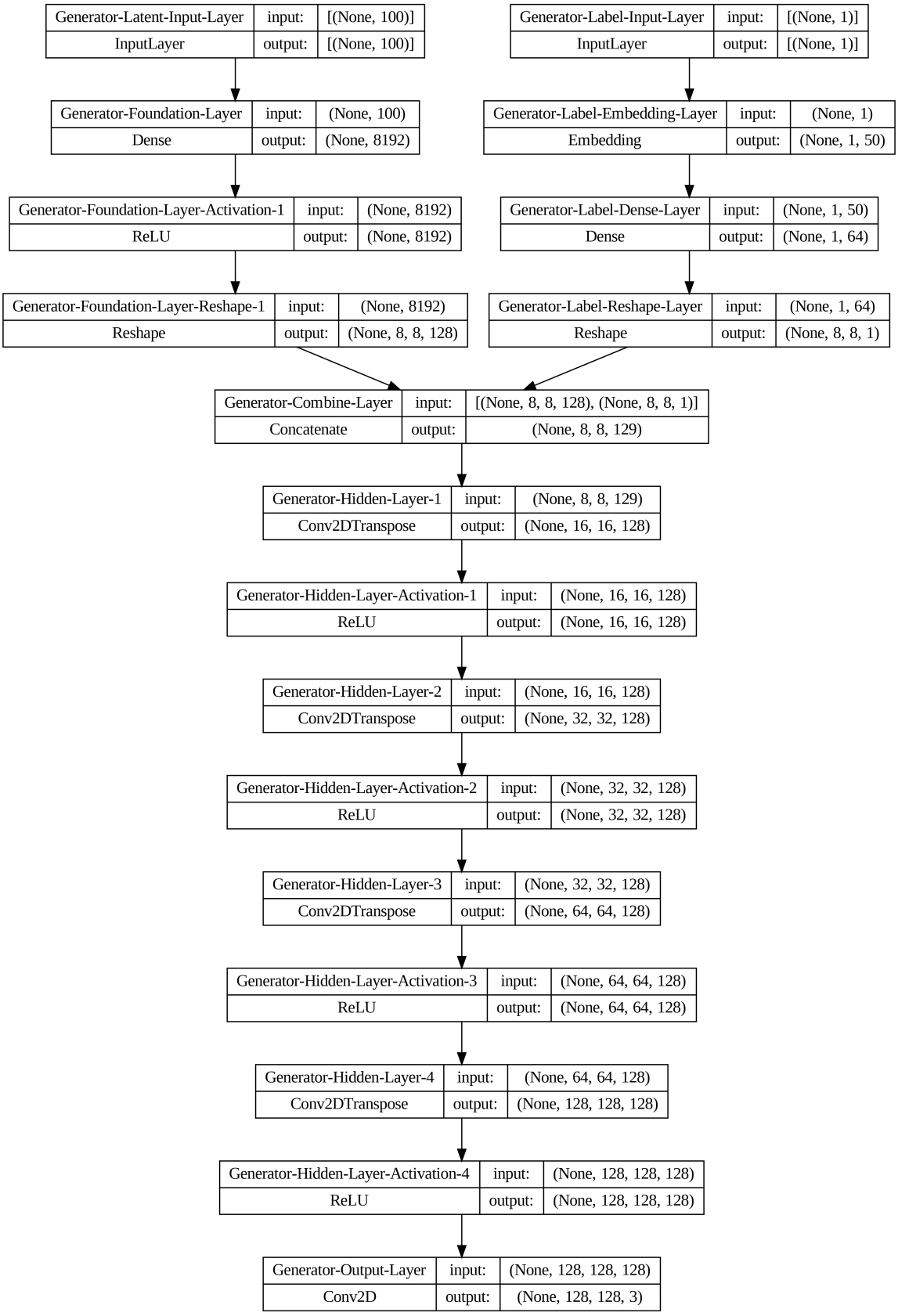

Figure 5.21: Network structure of the concordant cDCGAN generator

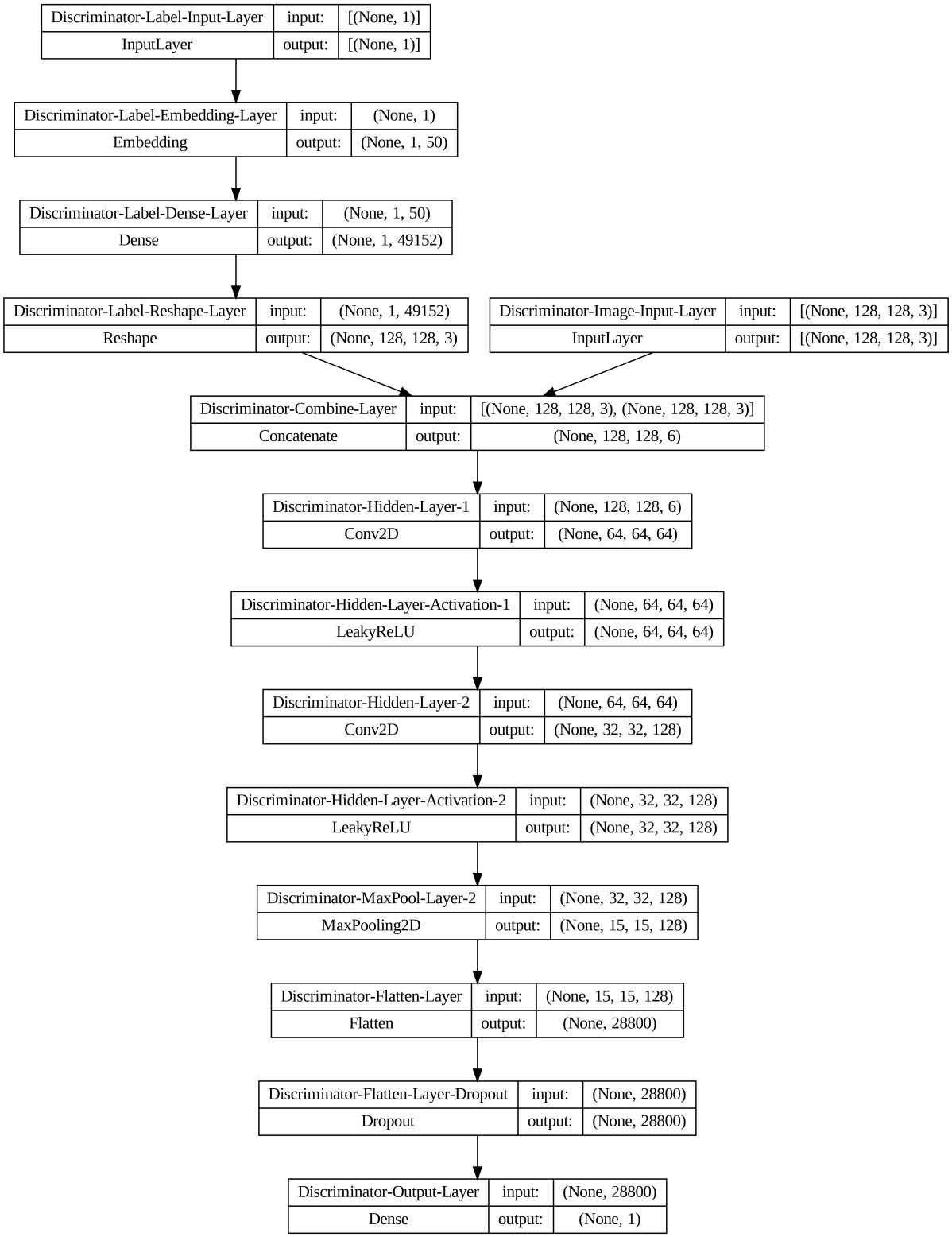

Figure 5.22: Network structure of the concordant cDCGAN discriminator

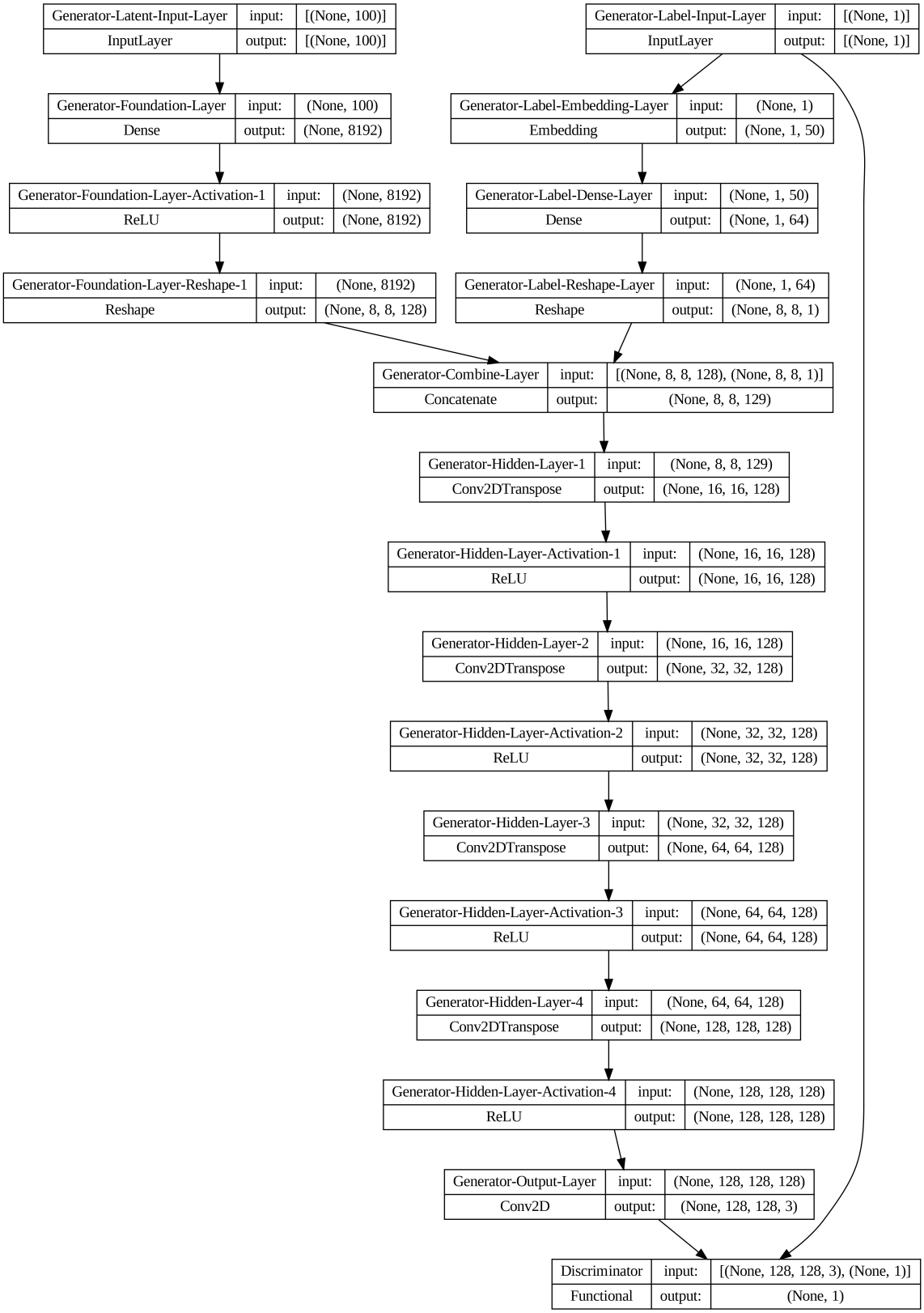

Figure 5.23: Network structure of the concordant cDCGAN generator and discriminator combined

Table 5.22: Confusion matrix of all datasets with and without concordant classifier integration

Dataset		Actual Label	Predicted Label		
			0	**1**	**2**
Non-Concordant Dataset	UCI	**0**	**450**	64	78
		1	70	**175**	53
		2	72	59	**350**
	PICCOLO	**0**	**460**	78	54
		1	50	**200**	48
		2	82	20	**379**
	CPDC	**0**	**2712**	265	-
		1	179	**1098**	-
Concordant Dataset	UCI	**0**	**545**	33	14
		1	18	**201**	73
		2	29	58	**355**
	PICCOLO	**0**	**580**	6	6
		1	7	**258**	33
		2	5	34	**442**
	CPDC	**0**	**2876**	101	-
		1	79	**1198**	-

5.2.1 FID, IS, PSNR and SSIM Analysis

The computed FID of our designed network is 4.5 for UCI, 4.9 for the PICCOLO, and 5.6 for the CPDC dataset; as the values obtained are less than 10 and closer to 0, it indicates that the generated images closely match the actual samples. Furthermore, the IS value achieved by the proposed model was 7.8, 8.2, and 7.5 on the UCI, PICCOLO, and CPDC datasets, respectively, showing that the generated images have good quality and diversity. In terms of PSNR, the mean values obtained were 20 dB, 22 dB, and 23 dB for three data sets, indicating acceptable image quality, and SSIM was 0.6 for UCI and 0.7 for PICCOLO and CPDC datasets, which shows a good similarity between original and synthetic images. The results of both metrics are shown in Table 5.21.

Table 5.23: Performance evaluation of updated UCI, PICCOLO and CPDC datasets with and without concordant classifier integration

Dataset		Label	Precision	Recall	F1 score
Non-Concordant Data	UCI	**0**	0.76	0.76	0.76
		1	0.59	0.59	0.59
		2	0.73	0.73	0.73
	PICCOLO	**0**	0.78	0.78	0.78
		1	0.67	0.67	0.67
		2	0.79	0.79	0.79
	CPDC	**0**	0.84	0.90	0.87
		1	0.78	0.68	0.73
Concordant Dataset	UCI	**0**	0.92	0.92	0.92
		1	0.69	0.69	0.69
		2	0.80	0.80	0.80
	PICCOLO	**0**	0.98	0.98	0.98
		1	0.87	0.87	0.87
		2	0.92	0.92	0.92
	CPDC	**0**	0.97	0.97	0.97
		1	0.94	0.92	0.93

5.2.2 Performance Evaluation

A deep concordant classifier was developed to effectively deal with the lack of data along with the complex morphology of colorectal polyps. Data insufficiency is a significant obstacle in training an effective CRC classifier. Therefore, for effective data augmentation necessary to perform improved classification, cDCGAN unified with ViT was developed. The confusion matrix obtained for each class is shown in Table 5.22. Results show that the capability of the classifier to identify instances of every class correctly has improved by segregating the images for model training.

Moreover, Table 5.23 presents the performance evaluation for individual classes of all datasets. The evaluation of advanced metrics is presented in Table 5.24, showing

Table 5.24: Performance evaluation of updated UCI, PICCOLO, and CPDC datasets with and without concordant classifier integration

Dataset		Accuracy	Precision	Recall	F1 score	Micro F1 score	Macro F1 score	Weighted F1 score	Cohens kappa Coefficient
Non-Concordant Dataset	UCI	0.74	0.69	0.67	0.68	0.72	0.69	0.71	0.5
	PICCOLO	0.76	0.73	0.75	0.74	0.76	0.75	0.76	0.6
	CPDC	0.83	0.86	0.79	0.83	0.83	0.80	0.83	0.6
Concordant Dataset	UCI	0.97	0.97	0.97	0.97	0.89	0.90	0.93	0.6
	PICCOLO	0.93	0.92	0.89	0.90	0.83	0.93	0.93	0.7
	CPDC	0.96	0.95	0.95	0.95	0.96	0.95	0.96	0.9

that in the case of the non-concordant dataset, the achieved macro F1 score was 0.69, and the weighted F1 score was 0.71. Meanwhile, the analysis of the concordant dataset showed a significant increase to a 0.90 macro F1 score and 0.93 weighted F1 score. In the case of the PICCOLO dataset, the macro F1 score value moved up from 0.75 to 0.93, and the weighted F1 score was enhanced from 0.76 to 0.93. Moreover, Cohen's Kappa Coefficient for concordant datasets was higher than for non-concordant data. In the case of UCI, the kappa value increased from 0.55 to 0.64; for PICCOLO, the value improved from 0.62 to 0.72; and in the case of the CPDC dataset, the kappa value enhanced from 0.6 to 0.9, which indicates substantial agreement.

5.2.3 Specificity and Sensitivity Analysis

Sensitivity, or the true positive rate, is a key metric for evaluating the performance of the model. Sensitivity assesses the proportion of correctly identified positive instances. It is a crucial factor in medical diagnosis, as failing to identify a disease has serious

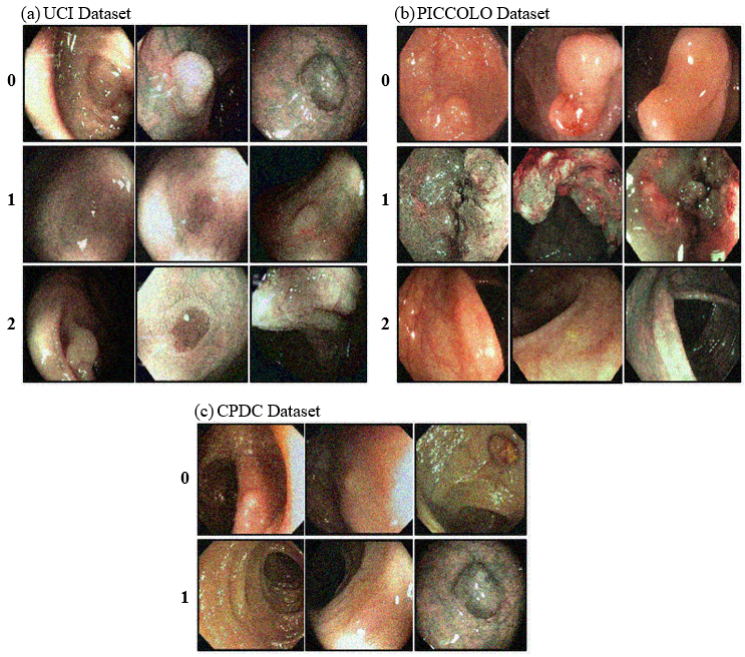

Figure 5.24: Concordant synthetic images of all datasets generated by the model

consequences. Experiments were performed on both non-concordant and concordant datasets, and analysis showed that the sensitivity value for the UCI dataset showed at least 3% improvement, and for the PICCOLO dataset, the sensitivity value was increased by a minimum of 5%.

In the medical domain, sensitivity and specificity are also beneficial in evaluating the classification model. It is a key metric to evaluate correctly identified negative instances. Specificity was evaluated for both non-concordant and concordant datasets. Experiments on UCI and PICCOLO showed 3% and 6% improvement in specificity values. Specificity and sensitivity are often considered together for a more comprehensive model evaluation. In a few cases, finding a trade-off between both metrics is desirable. As the given problem is a medical diagnosis, a highly sensitive model is crucial. Figure 5.25(a) presents that the proposed method improved the true positive and true negative

rate, i.e. the non-concordant dataset has less capability to identify the positive and negative instances correctly than the concordant dataset.

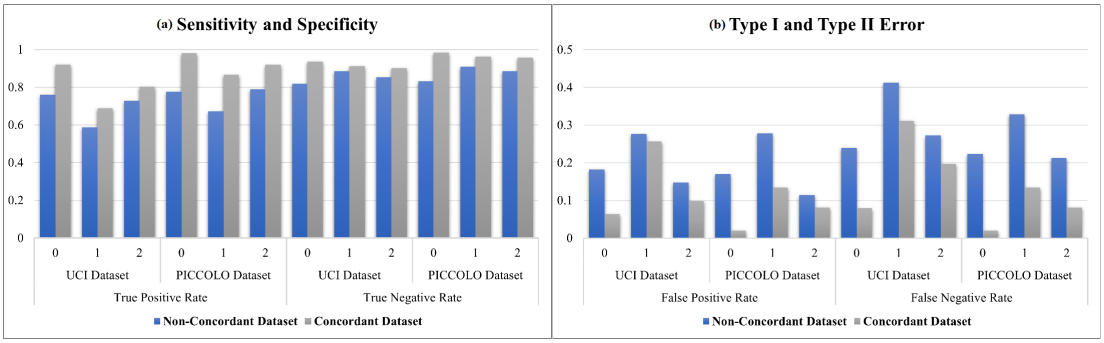

Figure 5.25: **(a)** Comparison of sensitivity, specificity, and **(b)** Type I, Type II error on UCI and PICCOLO dataset for non-concordant and concordant datasets

5.2.4 Type I and Type II Error Analysis

Type I error (False positive rate) can have a number of implications on patients' mental health. It can lead to unnecessary anxiety in patients about having a disease that does not exist, or it might lead to unnecessary diagnostic procedures. Therefore, it is also necessary that the classification model has a lower percentage of Type I error. The analysis of the proposed model presented in Figure 5.25(b) shows that the false positive rate was reduced up to 1% on both datasets. It was noted that the concordant dataset used for training the dataset improved the model results.

Type II error (false negative rate) occurs when the model fails to identify cancer in a patient that is present in the endoscopy image. It is crucial to handle type II errors because if a critical condition is not detected, it can risk a patient's life by delaying medical interventions and allowing the medical condition to progress further. Furthermore, an increased type II error reduces the sensitivity of the model by diminishing its ability

to identify true positive instances. The proposed model also focused on reducing type II error, and the results showed that it was reduced by 2% for both datasets.

Table 5.25: Performance comparison of proposed model

Dataset	Method	Accuracy	Precision	Recall	F1 score
UCI	Optimized CNN Network	92.7	0.91	0.92	0.91
	Ensemble Learning	93.6	0.93	0.94	0.93
	Weighted Ensemble DL Younas et al. (2023)	96.3	0.95	0.97	0.96
	Proposed Method	**97.5**	**0.97**	**0.97**	**0.96**
PICCOLO	Optimized CNN Network	73.2	0.73	0.72	0.73
	Ensemble Learning	77.3	0.79	0.74	0.77
	Weighted Ensemble DL Younas et al. (2023)	81.2	0.82	0.81	0.81
	Proposed Method	**93.3**	**0.92**	**0.89**	**0.90**

The results obtained using this method are compared with our previous study (Younas et al., 2023), which used CNN architectures on the same datasets with traditional data augmentation techniques. The comparison results shown in Tables 5.25 indicate that the proposed method gave significantly improved results. In the UCI and PICCOLO datasets, ViT further improved the results and attained 97.5% and 93.3% accuracy, 0.97 and 0.92 precision, 0.97 and 0.89 recall, and 0.96 and 0.90 F1 scores, respectively.

5.3 Model Explainability and Manifold regularisation

In the next part of the experiment, we extracted the features learned by the vision transformer network to build an explainable model. We trained our model using an explainable AI tool, Grad-CAM. This tool utilises the gradients of CRC polyp images, streaming them into the last layer of ViT to provide a coarse localisation indicating the

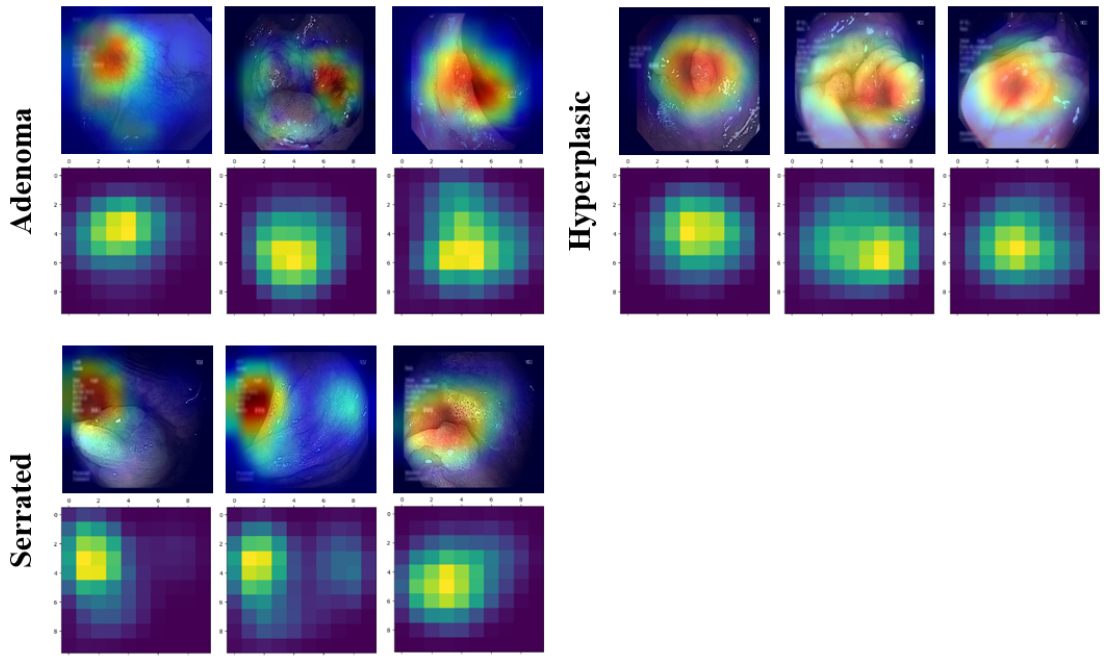

Figure 5.26: Generated explainable sample images for UCI dataset

critical areas of the images contributing to the decision-making of the model. The heat map is generated with an emphasis on the region for the predicted CRC polyp image. The highest gradient values are present in the regions where polyps exist in the image. The output of Grad-CAM provides the information that took part in generating the final decision. Moreover, superimposed images are also generated along with the heatmaps to highlight the regions of interest in the polyp images. Figures 5.26, 5.28, and 5.30 shows a few samples of the superimposed images and their corresponding heatmaps generated for the UCI, PICCOLO and CPDC datasets.

Once we had extracted the features using the vision transformer and verified the decision-making process, non-linear manifold learning algorithms were applied to the extracted features. The main purpose of this step is to make the process less computationally expensive by extracting a subset of the most important features and

regularising the model simultaneously by handling the overfitting problem. Several manifold learning algorithms were applied to each dataset separately to analyse the result. Regarding the UCI dataset, the standard and modified LLE gave the best results for ARI and NMI metrics. Among all the implemented manifold algorithms, the model achieved the highest values for LLE: 0.842, 0.832 NMI, and 0.781, 0.762 ARI; in addition, the FMI value achieved by modified LLE was also the highest of 0.862 in comparison to all the other algorithms. The result presented in Table 5.26 indicates that the model was able to reduce the dimensionality while preserving its geometry. Moreover, Figure 5.27 provides the visual representation of the non-linear dimensionality reduction results. A different colour indicates each class to make the visualisation easier.

Table 5.26: Comparison between various manifold learning algorithms for the optimum components for UCI dataset

Dataset	**Backbone**	**Method**		**ARI**	**NMI**	**FMI**
UCI	cDCGAN + ViT	Isomap		0.744	0.805	0.613
		t-SNE		0.707	0.779	0.555
		LLE	Standard	0.781	0.842	0.681
			Modified	0.762	0.832	0.862
			Hessian	0.765	0.794	0.585
			Tangent Space Alignment	0.734	0.717	0.624
		Truncated SVD		0.699	0.671	0.672
		Spectral Embedding		0.654	0.697	0.608
		Random Trees		0.765	0.797	0.697
		LDA		0.717	0.734	0.652

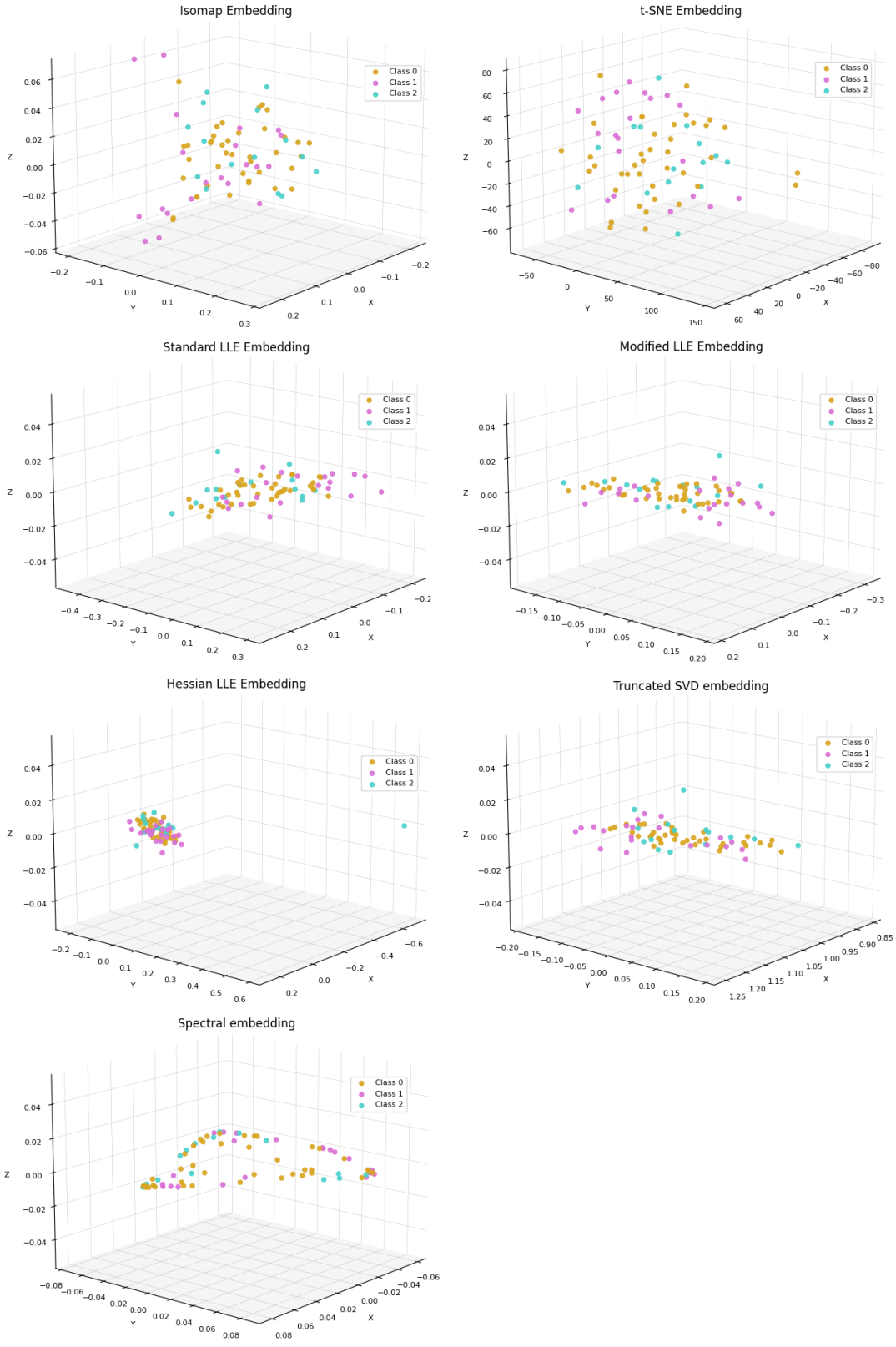

Figure 5.27: Manifold learning visualizations on UCI dataset

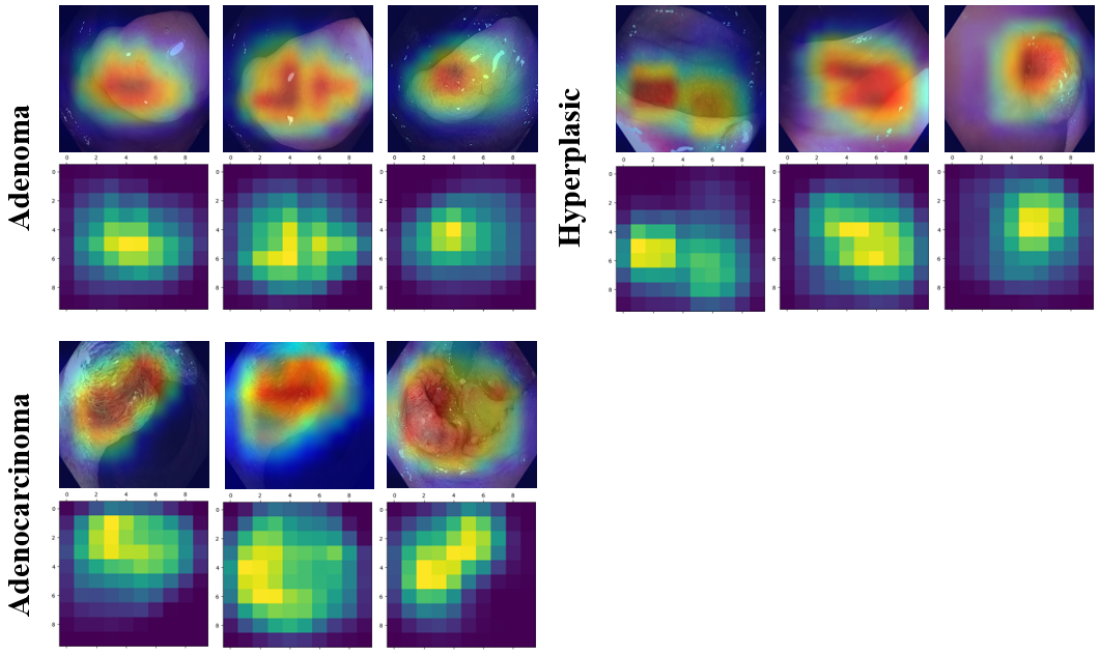

Figure 5.28: Generated Explainable Sample Images for PICCOLO Dataset

The same procedure was performed on the PICCOLO dataset, and the obtained results are shown in Table 5.27. Analysis of results shows that the LLE provides a consistent performance on this dataset as well. However, instead of Modified LLE, Standard LLE achieved the highest results of 0.887 ARI, 0.904 NMI and 0.941 FMI. The second best-performing algorithm is still the Modified LLE, which has the same performance as the UCI dataset; it achieved 0.878 ARI, 0.868 NMI, and 0.773 FMI values. Figure 5.29 provides a visual representation of the non-linear dimensionality reduction of extracted features from the PICCOLO dataset.

Table 5.27: Comparison between various manifold learning algorithms for the optimum components for PICCOLO dataset

Dataset	Backbone	Method		ARI	NMI	FMI
PICCOLO	cDCGAN + ViT	Isomap		0.773	0.805	0.722
		t-SNE		0.717	0.779	0.647
		LLE	Standard	0.887	0.904	0.941
			Modified	0.878	0.868	0.828
			Hessian	0.786	0.815	0.773
			Tangent Space Alignment	0.784	0.805	0.728
		Truncated SVD		0.655	0.611	0.609
		Spectral Embedding		0.644	0.457	0.531
		Random Trees		0.413	0.675	0.617
		LDA		0.716	0.739	0.694

Lastly, the results of CPDC datasets were examined, and the results are illustrated in Table 5.28. The CPDC dataset consists of two classes, and the experiment results are somewhat different from those of the previously employed three-class datasets. We observed that the best results were achieved through Modified LLE again, producing 0.871 ARI, 0.58 NMI and 0.923 FMI; however, the standard and hessian LLE did not perform well for this dataset. Moreover, in this scenario, the Isomap algorithm also performed well, generating the second-best result of 0.769 ARI, 0.861 NMI, and 0.771 FMI values. Figure 5.31 shows the outcome of manifold learning algorithms on the CPDC dataset.

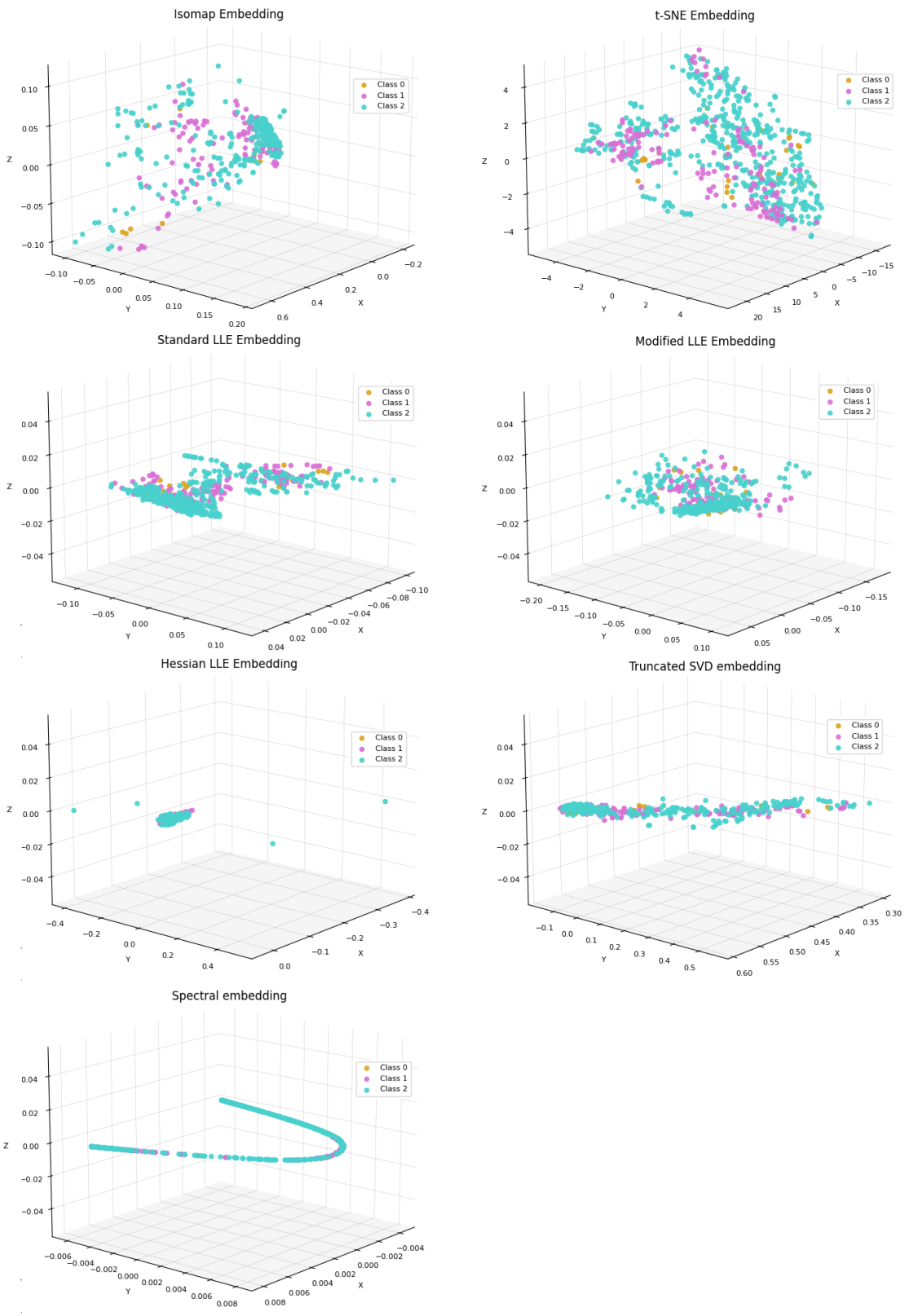

Figure 5.29: Manifold learning visualizations on PICCOLO dataset

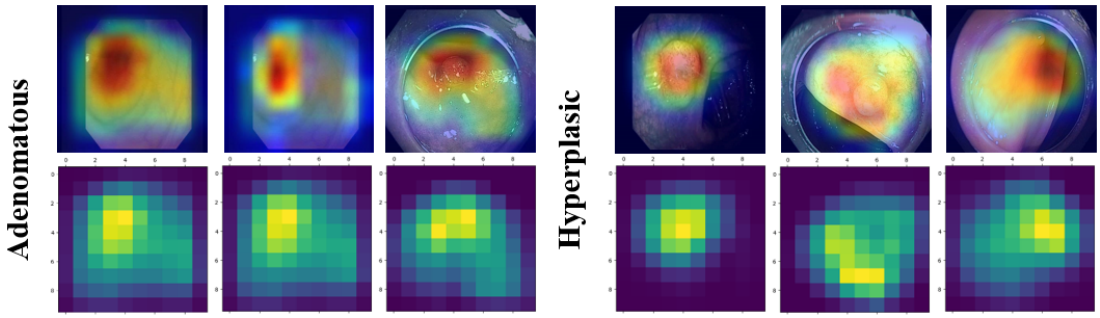

Figure 5.30: Generated explainable sample images for CPDC dataset

Table 5.28: Comparison between various manifold learning algorithms for the optimum components for CPDC dataset

Dataset	Backbone	Method		ARI	NMI	FMI
CPDC	cDCGAN + ViT	Isomap		0.769	0.861	0.771
		t-SNE		0.601	0.623	0.591
		LLE	Standard	0.670	0.789	0.633
			Modified	0.871	0.858	0.923
			Hessian	0.637	0.599	0.613
			Tangent Space Alignment	0.671	0.699	0.601
		Truncated SVD		0.689	0.760	0.728
		Spectral Embedding		0.700	0.757	0.695
		Random Trees		0.483	0.707	0.573
		LDA		0.601	0.637	0.591

The comparison of ARI, NMI and FMI metrics exemplifies that the LLE consistently gives better results on all datasets, indicating superior clustering performance among the tested models. Figures 5.32, 5.33, and 5.34 presents the comparison of each

metric. The high performance of LLE can be associated with its efficacy in preserving local geometry, which is imperative for clustering.

The weight matrices of LLE were built on the assumption that every sample in the dataset is a linear function of two neighbouring data points. It was found during the experimentation that a non-linear manifold works more efficiently with a lower number of neighbours. Moreover, to preserve the original classifier loss, the tuning parameters α, β and γ were initialized to 10, 0.1 and 0.0001. These settings ensure that the classification loss is not overridden by the non-linear manifold learning, and CNN parameters carry out the learning.

5.4 Concluding Classifier

In the final stage, the model regularisation block results were passed on to the CNN layers for final classification. The confusion matrices are presented in Figure 5.35, and the performance evaluation for individual classes is shown in Table 5.29. Results show that the designed framework can distinguish between the polyp classes effectively. Moreover, the overall performance results obtained are shown in Table 5.30 and the evaluation of advanced metrics is given in Table 5.31.

The obtained results indicate that the developed framework produced impressively accurate results with minute errors on all the datasets. Among the three datasets, CPDC provided the best results, being a two-class problem, whereas UCI and PICCOLO also generated noticeable results, considering all the challenges posed by these datasets. As mentioned earlier, sensitivity or the true positive rate is a key metric for evaluating the performance of the model, and it is a crucial factor in medical diagnosis, as failing

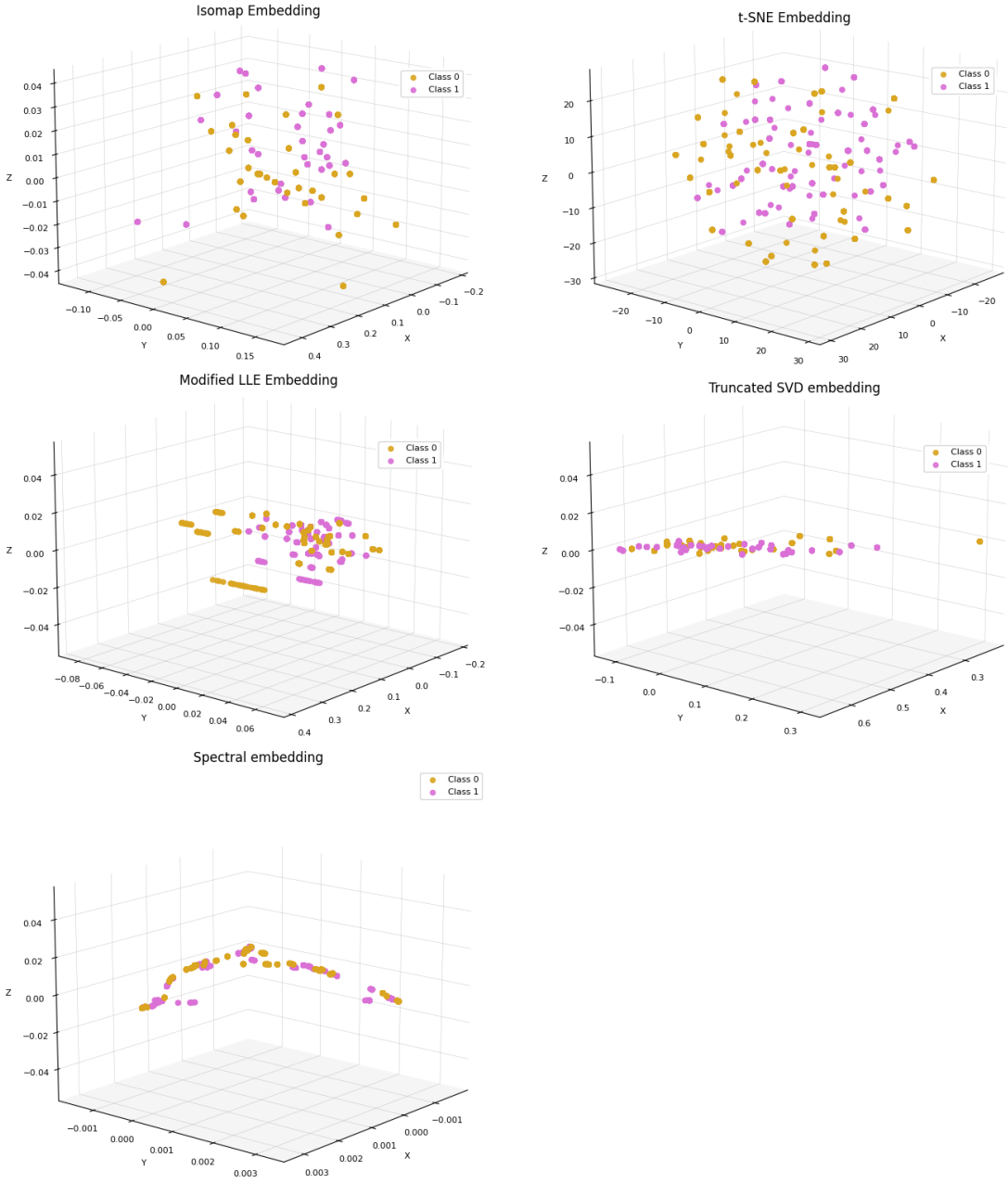

Figure 5.31: Manifold learning visualizations on CPDC dataset

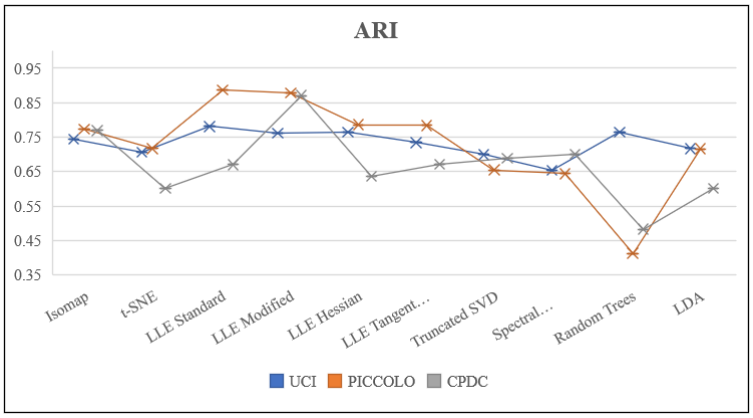

Figure 5.32: ARI comparative visualization on UCI, PICCOLO, and CPDC dataset

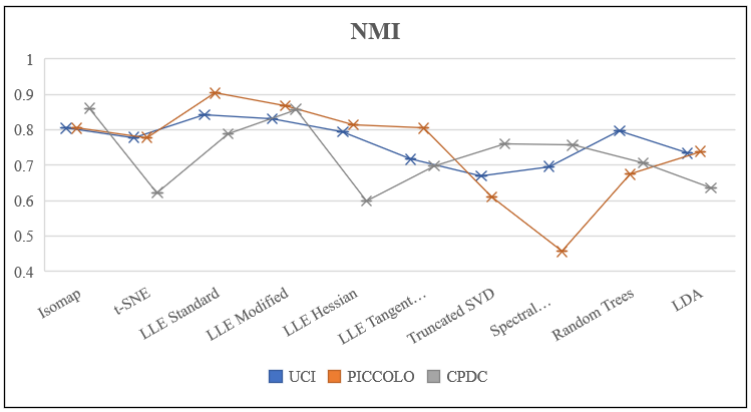

Figure 5.33: NMI comparative visualization on UCI, PICCOLO, and CPDC dataset

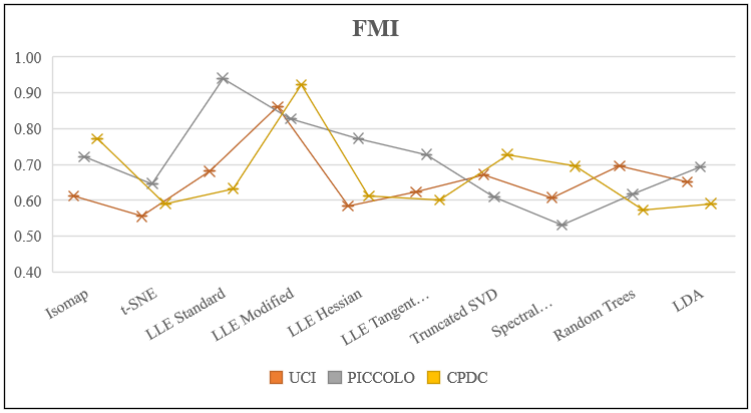

Figure 5.34: FMI comparative visualization on UCI, PICCOLO, and CPDC dataset

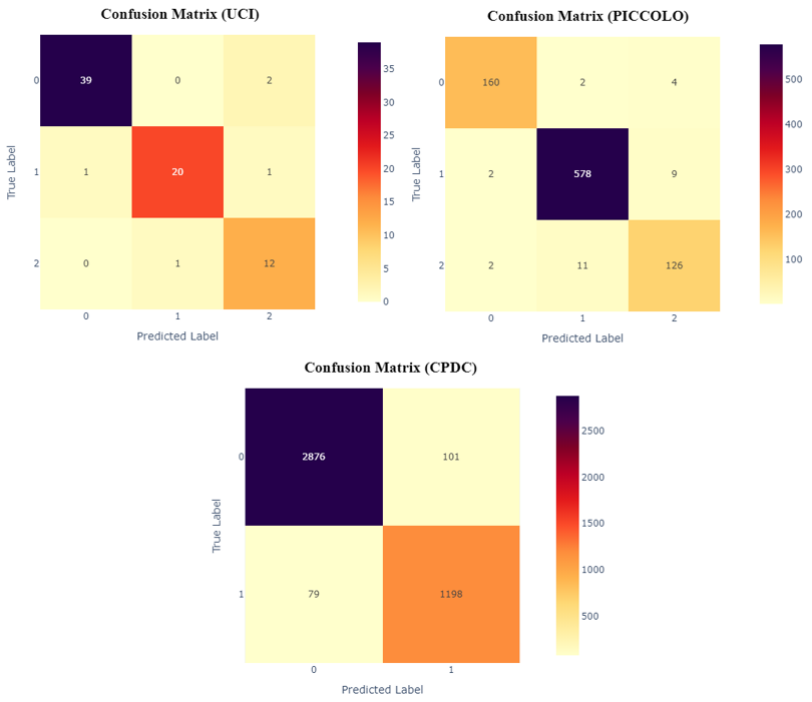

Figure 5.35: Confusion matrix visualization of UCI, PICCOLO, and CPDC dataset

to identify a disease has serious consequences. Experiments were performed on all the datasets, and analysis of the concluding classifier showed that the sensitivity values for all datasets were above 95%. Like sensitivity, specificity is also beneficial for evaluating the classification model. It is a key metric to evaluate correctly identified negative instances. The experiments showed that the specificity values range from 92% to 97%, indicating that the model has a high capacity to effectively classify positive and negative instances.

Table 5.29: Confusion matrix and performance evaluation on all validation datasets for proposed classification model

Dataset		Predicted Label					
	Label	0	1	2	**Precision**	**Recall**	**F1-score**
UCI	0	**39**	0	2	0.98	0.95	0.96
	1	1	**20**	1	0.95	0.91	0.93
	2	0	1	**12**	0.80	0.92	0.83
PICCOLO	0	**160**	2	4	0.96	0.96	0.96
	1	5	**578**	9	0.98	0.98	0.98
	2	2	11	**126**	0.91	0.91	0.91
CPDC	0	**2876**	101	-	0.97	0.97	0.97
	1	79	**1198**	-	0.94	0.92	0.93

Table 5.30: Performance measures of proposed model for all datasets

Dataset	**Precision**	**Recall**	**F1 score**	**Accuracy**	**Sensitivity**	**Specificity**
UCI	0.92	0.95	0.93	0.93	0.95	0.92
PICCOLO	0.97	0.96	0.97	0.96	0.96	0.96
CPDC	0.96	0.95	0.96	0.96	0.96	0.95

All classes have data imbalance, and two of the datasets require multiclass classification that requires results evaluation through advanced metrics. The macro f1 score was 0.93, 0.96 and 0.97 for the UCI, PICCOLO and CPDC datasets. MCC and Kappa values were also evaluated for the final classification task. MCC values obtained range from 0.90 to 0.93, and the kappa values showed a significant improvement compared to the results of the previous methodologies. The kappa values show that the model has a high degree of separability.

Table 5.31: Advanced metric evaluation of proposed model for all datasets

Dataset	Micro F1 score	Macro F1 score	Weighted F1 score	Error	MCC	Kappa
UCI	0.93	0.92	0.93	0.07	0.90	0.6
PICCOLO	0.96	0.95	0.96	0.04	0.91	0.6
CPDC	0.96	0.95	0.96	0.04	0.93	0.9

5.5 Qualitative Analysis: Model Decision Making in Challenging Cases

- **Small or Flat Polyps with Low Contrast**

  Small or flat polyps often blend with the surrounding mucosa, making them difficult to detect due to low contrast and ambiguous boundaries. Traditional CNNs struggle in these cases because they rely on local feature extraction, which may not capture subtle intensity variations effectively. The ViT architecture captures global dependencies between patches, enabling it to identify subtle contrast variations that define polyp boundaries. Grad-CAM heat maps show strong activation in polyp regions, even when CNN-based models do not highlight them, demonstrating the model's ability to distinguish small polyps from normal mucosa. In some borderline cases, the model still misclassifies polyps when their contrast difference is negligible, indicating a need for contrast enhancement techniques to improve sensitivity.

- **Poorly Illuminated or Noisy Endoscopy Frames**

  Shadows, reflections, and motion blur in endoscopic frames degrade image quality,

leading to false negatives or misclassifications. CNNs typically struggle with these artefacts due to their local receptive fields and limited robustness to variations in illumination. The ViT model compensates for noise better than CNNs because its self-attention mechanism processes long-range dependencies, allowing it to infer missing or distorted features from surrounding image regions. Performance is notably better in low light conditions, where CNNs fail due to missing texture details. However, extreme motion blur leads to misclassification, as the model cannot reliably extract discriminative features. This highlights the need for frame selection algorithms that filter out severely degraded frames before processing.

- **Rare Polyp Morphologies**

  Some polyp types, such as serrated adenomas, exhibit hybrid structural characteristics that may not be well represented in the training dataset. This results in lower classification accuracy for these rare cases. To address data scarcity, CGAN-based augmentation introduces synthetic rare polyp samples, which improves the model's generalisation to under represented classes. Grad-CAM analysis reveals occasional attention misalignment, where the model mistakenly prioritises surrounding mucosa instead of the polyp itself, leading to false negatives. This suggests a need for fine-tuning with additional hard-negative mining to force the model to learn subtle but clinically significant features of rare polyps.

5.6 Summary of Chapter

This chapter presents the comprehensive outcomes of applying the designed framework to the CRC polyp classification task and provides an in-depth discussion of the findings.

The analysis is structured to evaluate each framework block individually based on relevant performance metrics. This thorough evaluation ensures that each component's contribution to the overall performance is well understood and documented.

The first part of this chapter delves into the results obtained from extensive experimentation with the ensemble learning blocks. This section discusses the detailed findings from hyperparameter tuning, transfer learning, and the application of ensemble learning techniques. The impact of various hyperparameters on model performance is analysed to determine which configurations yield the best results. Hyperparameter tuning, including adjustments to learning rates, batch sizes, and epoch numbers, is crucial to optimising model performance and achieving the desired accuracy. In addition, the effectiveness of transfer learning is examined. Transfer learning involves utilising pre-trained models and adapting them to the CRC polyp classification task. This approach leverages existing knowledge from other domains to improve model accuracy and reduce training time. The chapter discusses how transfer learning enhances the model's ability to generalise and perform well on the CRC polyp dataset, showcasing its benefits in improving classification performance.

In addition, the benefits of ensemble learning are explored in this section. Ensemble learning combines the strengths of multiple models to create a more robust and accurate overall system. The chapter explains how integrating various models through ensemble methods enhances the performance and robustness of the classification system. Highlights how ensemble learning improves the model's ability to handle complex and diverse data, leading to more reliable and accurate predictions.

To evaluate the image generation block, its performance is evaluated using

subjective measures such as the Fréchet Inception Distance (FID), Inception Score (IS), Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity Index(SSIM). FID compares the distribution of the generated images and the real images in terms of feature space, thus assessing the quality and realism of the images produced. On the other hand, IS assesses the degree of diversity and quality of the images generated based on their conformity to the target class distributions in the datasets. PSNR is a widely used metric for assessing the quality of a reconstructed or compressed image compared to its original counterpart. SSIM measures the perceptual similarity between two images. Such assessments provide a measurable evaluation of the efficiency of the synthetic image generation procedure in terms of its generative nature and also emphasise the enhancement of the synthetic generation process in terms of its quality and diversity.

The performance of the deep feature extraction block is assessed in other areas. This entails assessing the contribution of the deep features extracted from the images towards the classification of the images. Fixing the accuracy of the feature extraction process is important in the construction of a strong classifier since it guarantees that the most discriminative and applicable features are obtained and utilised. The evaluation includes how efficiently the features derived from both real and synthetic images form a basis for improving the performance of the model in the classification of CRC polyps. Furthermore, the understanding of the decision-making process of the model is analysed through various explainability techniques, which provide an insight into the workings of the prediction. Explainability functions such as AlexNet and grad-CAM are performed to visualise image-based data and determine which specific aspects of the images model leaned towards predicting which class. This is useful in the evaluation of the model's cognitive ability and whether the outcome achieved is what is expected with regard to

the features learnt.

In addition, the chapter further examines model performance with respect to the application of non-linear dimensionality reduction techniques and regularisation. Most of the time, high-dimensional data are resolved using nonlinear dimensionality reduction techniques like Isometric Mapping (Isomap) and Locally Linear Embedding (LLE) in order to enhance the performance of the model. Apart from the foregoing, regularisation methods are implemented to curb overfitting where increasing model complexity is required due to complex training data. The efficiency of these strategies is evaluated with the help of metrics including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). These measures are shown to improve the model's non-linear dimensionality reduction and regularisation robustness and interpretability framework.

Finally, we present the classification results and provide a detailed analysis of the predictive performance of the proposed framework. The evaluation includes a comprehensive discussion of performance metrics, including specificity, sensitivity, accuracy, precision, recall, F1-score, micro and macro F1-scores, Matthews Correlation Coefficient (MCC), and Cohen's Kappa values. Specificity measures the model's ability to correctly identify negative cases, while sensitivity (recall) evaluates its effectiveness in detecting true positives. Accuracy reflects the overall performance by combining correct classifications of positive and negative cases, whereas precision gauges the proportion of true positives among all positive predictions. The F1 score balances precision and recall, providing a metric that accounts for false positives and false negatives, with micro and macro F1-scores offering insights into performance across multiple classes. MCC provides a balanced measure of the performance of the model, particularly useful for

imbalanced datasets, and Cohen's Kappa evaluates the agreement between the model predictions and the actual classes, adjusting for chance. The classification results are compared with existing state-of-the-art deep learning approaches, highlighting the superior performance of the proposed framework. This comparison demonstrates that the proposed methodology not only meets but often exceeds the performance of current leading methods, showcasing its effectiveness in delivering accurate, reliable, and interpretable classification results for CRC polyp detection.

Chapter 6

Conclusion

This chapter concludes the research done in this thesis, reflects on the implication of the findings in the CAD domain, acknowledges the limitations of the research performed, and suggests future directions that we consider promising in the context of this work. This chapter aims to illustrate how the research questions and objectives outlined in Chapter 1 have been addressed and reflect the overview of the contribution of this study. Section 6.1 summarises the essential insights and elaborates on all the key findings of the research questions. Section 6.2 illustrates the theoretical and practical implications of the research. Finally, Section 6.3 expounds on the methodological limitations and scope of the study and highlights future research directions.

In recent years, the practice of artificial intelligence (AI) and deep learning (DL) technologies has continued to gain a lot of applications, and the field of medicine seems to be one of the greatest beneficiaries. Incorporating such advanced technologies into the field of medicine has made it possible to develop new techniques for diagnosis based on images, thus improving the accuracy and agility to predict the presence of a disease. Regarding colorectal cancer screening strategies (CRC), colonoscopy remains the cornerstone procedure for the identification and prophylaxis of CRC, being one of the most powerful diagnostic techniques.

Although this procedure is crucial, the practice of colonoscopy remains highly dependent on the ability, experience, and focus of the endoscopist performing this exact procedure. Colonoscopic analysis is by nature a very subjective examination procedure, which implies that the outcomes are subject to variations around the skill level of each practitioner. This is a major drawback since inconsistency in colonoscopy results can cause complications, such as missed or delayed diagnosis, especially colorectal cancer, which is often easier to treat when diagnosed early. Hence, the need for quality assurance of the procedures performed in colonoscopy is fundamental in the scope of strategies that aim to lower CRC rates in the community.

To tackle these issues, medical organisations have been actively working on different ways to improve the quality control processes of attention over call procedural examination. Earlier approaches included training and certification of endoscopists, formulating uniform guidelines, and employing several adjuncts during polyp surgery and other abnormalities. However, despite some of these improvements, the most prominent drawback remains the subjective view of the test. From this perspective, it can be said that deep learning-assisted colonoscopy or virtual biopsy could be considered an

exceptional supplementary method. These systems assist endoscopists in detecting subtle changes through spatial intelligence and deep learning techniques, enabling real-time assistance to the endoscopists during the procedure. The incorporation of deep learning models in performing colonoscopies will help reduce variability in human judgment and errors and promote uniform accuracy and quality of work in CRC screening. One area where great advancements have been made is the virtual biopsy where no cutting of any tissue is done as with the conventional approach of a tissue biopsy.

It can be asserted that these AI-assisted tools have been developed to improve the adequacy and consistency of the diagnosis of colorectal cancer in primary prevention. Since this area is also developing, deep learning for colonoscopy will likely become part of every CRC screening program and provide great support in the early detection and treatment of the disease. This change in attitude in terms of acceptance of AI in diagnostic medicine highlights the fact that technology can change the way medicine is practiced by making it more accurate and affordable to more people.

The research presented in this thesis was driven by a critical observation: the urgent need for a virtual biopsy tool for colorectal cancer (CRC) is essential not only to mitigate the risk of cancer-related deaths but also to alleviate the heavy workload faced by pathologists. While numerous techniques are available that can detect polyps effectively, there is a notable absence of a comprehensive framework capable of addressing the multitude of challenges posed by this complex medical dataset. The development of such a framework is essential, but it is fraught with various challenges, especially when multiple issues must be tackled simultaneously.

One of the primary challenges in developing computer-aided diagnostic (CAD)

systems is the availability and accessibility of medical datasets. These datasets are often scarce, incomplete, or difficult to obtain, which poses a significant barrier to creating effective diagnostic tools. Traditional data augmentation techniques, commonly used to address the lack of data, can inadvertently affect the feature extraction process negatively. These techniques often introduce redundancy and do not necessarily increase the diversity of the dataset, which is crucial for training robust models. At the same time, extracting important features that include fine-grain details is critical to reducing the misclassification rate of polyps. However, focusing on this aspect introduces the problem of high dimensionality in the data, which complicates the development of a robust and generalised model.

There is a high dimensionality in every data source that can pose a challenge, especially in pattern recognition in various datasets, which causes overfitting. The model learns the training data so accurately that when presented with data for which it was not trained on, it does not perform well. To counter this, dimensionality reduction techniques such as PCA and MDS attempts to eliminate some extraneous information. Yet, these techniques are linear and are unable to reveal the true nature of high-dimensional data, leading to almost no meaningful improvement to the performance of the model.

Additionally, besides the technical hurdles, a major barrier to the successful adoption of these CAD systems in everyday clinical practice is the pathologists' lack of confidence in deep learning solutions' involvement in the decision-making process. The "black-box" nature of the majority of the current deep learning technologies reduces the clarity of decision-making in most health-related issues, such as how a clinician reaches a particular diagnosis or offers a prognosis. This lack of clarity is troubling because pathologists do not want to work with a system they do not understand because it can

affect a decision relating to the welfare of a patient.

Given these numerous challenges, it becomes clear that a suitable and integrated methodology is required to handle all these problems simultaneously and seamlessly. The methodology must address the issues of data availability, feature extraction, high dimensionality, overfitting, and transparency in the decision-making process. The research aims to develop a comprehensive framework that incorporates advanced techniques such as non-linear dimensionality reduction, robust feature extraction methods, and transparent deep learning models to create a reliable and trustworthy virtual biopsy tool for CRC. This approach aims to improve CRC detection and classification accuracy and reliability. Also, it seeks to build a bridge of trust between advanced AI-driven systems and the medical professionals who use them. By holistically addressing these challenges, this research sets the foundation for successfully implementing CAD systems in clinical practice, ultimately contributing to better patient outcomes and more efficient healthcare delivery.

The development of a well-designed structure for detection and classification is crucial for colorectal cancer (CRC), which can not be accomplished by the integration of already available technologies in a linear manner and expects positive results. Every phase in this undertaking has to be given proper and efficient attention to detail so as to ensure the effectiveness and correctness of the end product is achieved. The process starts with data cleansing, which comes first in setting up such a framework, where the role played is very critical for every other step to follow. This stage is not merely an initial stage but rather an important one which affects all other processes from data augmentation up to the last stage of carrying out classification. For example, data augmentation goes hand in hand with care to avoid new redundancy or noise and increase

only the quality and diversity of the dataset. This justification is crucial and provides the basis of why the next step is necessary, as it is beneficial towards enhancing the generalisation performance of the model on unseen data. The following stage deals with feature extraction, which is directed to obtain and capture important features which can help bring down the misclassification rate of polyps. On the other hand, this step brings about the problem of working with a high dimensional feature vector, which can result in overfitting if no measures are taken. To tackle the issue of high dimensionality, the framework employs advanced dimensionality reduction techniques. These techniques are not just applied superficially but are the result of extensive experimentation and analysis to determine the most effective methods for the specific dataset at hand. The goal is to reduce the data dimensionality while preserving its intrinsic structure, thereby enhancing the model's ability to learn meaningful patterns without being overwhelmed by the complexity of the data.

Model regularisation is another critical aspect of the framework that requires careful consideration. This step involves fine-tuning the model to prevent overfitting, ensuring it remains robust and generalisable across different datasets. The regularisation process is not a one-size-fits-all solution but is tailored to the specific needs of the framework, based on the characteristics of the data and the requirements of the task at hand. The training phase of the developed method is particularly resource-intensive, requiring significant computational power due to its separate training phases for base classifiers, intensive image synthesis, and complex feature extraction processes. The framework's complexity is further compounded by the need for elaborate experimentation on dimensionality reduction and model regularization, all of which contribute to the overall effectiveness of the final model. Despite these challenges, using advanced

GPUs helps mitigate the computational burden, allowing for efficient processing of large datasets and complex models.

Moreover, adopting advanced, customised methodologies proves to be adequate and essential for managing the complex tasks involved in CRC detection and classification. These methodologies are designed to address the specific challenges posed by the medical dataset, such as class imbalance, high dimensionality, and the need for explainable models. Integrating these methodologies into a cohesive framework ensures that each step is optimised for the best possible outcome, leading to a more accurate and reliable diagnostic tool. Therefore, developing a robust CRC detection and classification framework requires a careful and intelligent approach at every stage of the process. From data pre-processing to final classification, each step is meticulously planned and executed to overcome the various challenges inherent in working with medical datasets. The resource-intensive nature of the training phase is managed through the use of GPUs. At the same time, the adoption of advanced, customised methodologies ensures that the framework is capable of handling the complexities of the task. This integrated approach not only improves the accuracy and reliability of the final model but also lays the groundwork for the successful implementation of computer-aided diagnostic systems in clinical practice.

As previously discussed, developing an effective computer-aided colorectal polyp classifier necessitates a large dataset to achieve successful image classification. However, the benchmark datasets available for this project were insufficient to meet this need. To address the issue of data insufficiency, we conducted a comprehensive exploration focused on building a sensitive classifier, with a primary emphasis on enhancing precision in cancer classification. The proposed model exhibited exceptional capabilities

in classifying colorectal polyp images, tackling the challenges posed by limited data.

To overcome the limitations of the available datasets, we employed cDCGAN to generate synthetic images. However, the generated data was a mix of good and poor-quality images, which had the potential to negatively impact the classifier's performance. To mitigate this issue, the proposed framework implemented a mechanism to segregate good-quality images from poor-quality ones based on class probability. This segregation ensures that only high-quality synthetic images contribute to the training process, thereby enhancing the overall accuracy of the model. To further improve feature extraction, we integrated ViT with cDCGAN. ViT, a recent advancement in pure transformers, has shown promising results in the computer vision domain, making it an ideal choice for this task. The use of ViT allowed for the extraction of fine-grain features that were crucial for accurate classification. The incorporation of these features also facilitated the integration of XAI into the designed framework, enhancing the transparency and interpretability of the model's decisions. Additionally, manifold learning played a pivotal role in dimensionality reduction and model regularisation, addressing the challenges associated with high-dimensional data. By preserving the intrinsic structure of the data, manifold learning ensured that the model could generalise well, reducing the risk of overfitting and improving its robustness. This combination of advanced techniques resulted in a framework that effectively handles the complexities of CRC polyp images.

Our experimental results demonstrated that the Vision Transformer model, when adapted and combined with a concordant conditional adversarial network for CRC image classification, outperformed traditional CNNs trained on data augmented through conventional methods. The superior performance of the proposed method was evident through increased accuracy, reaffirming its effectiveness in managing the intricacies

associated with CRC polyp images.

The effectiveness of the proposed model was further validated by the values obtained for various evaluation metrics. The FID, IS, PSNR and SSIM metrics, used to assess the quality and diversity of synthetic image data, confirmed that the synthetic images generated by the cDCGAN were of high quality and contributed positively to the classification process. Moreover, ARI, NMI, and FMI metrics, which were used to evaluate the effectiveness of manifold learning, indicated that the proposed model efficiently handled the challenges of dimensionality reduction and model regularisation.

The proposed framework, which integrates cDCGAN, ViT, and manifold learning, offers a robust solution for CRC image classification. It not only addresses the issue of data insufficiency but also enhances the precision and reliability of the classification process. The comprehensive evaluation of the model using various metrics underscores its potential as a valuable tool in the field of medical image analysis, particularly for colorectal cancer diagnosis.

A comparative analysis with our previous study on the same datasets highlights the significant superiority of the proposed method. The comparison was meticulously conducted using both non-concordant and concordant versions of the UCI, PICCOLO, and CPDC datasets. The results of this comparison clearly demonstrate that the proposed method excels in classifying polyps with notably higher accuracy. This improvement in accuracy is not merely incremental; it marks a substantial advancement in the reliability and effectiveness of polyp classification. Beyond its impressive classification accuracy, what truly distinguishes our method is its heightened sensitivity, which is a critical factor in medical diagnosis. Sensitivity, in this context, ensures that the model can accurately

identify malignant polyps, thereby minimising the risk of false negatives, an outcome that could have severe consequences in a clinical setting. The increased sensitivity of our model thus enhances its utility as a diagnostic tool, offering pathologists a more reliable and trustworthy resource.

The findings presented in this thesis make a meaningful contribution to the broader discourse on the transformative impact of generating concordant images and extracting fine-grain feature details. These elements are not just technical improvements; they are foundational steps towards building an explainable model that can be confidently implemented in clinical settings. The ability to explain the model's decisions is crucial for gaining the trust of medical professionals and ensuring the adoption of such tools in real-world practice.

Moreover, the study underscores the potential of the proposed method in advancing the field of precision medicine. Precision medicine aims to tailor medical treatment to the individual characteristics of each patient, and the accuracy, sensitivity, and explainability of our method align perfectly with these goals. By offering a tool that not only performs well in terms of classification accuracy but also provides clear insights into the decision-making process, this research paves the way for more personalised and effective medical interventions.

The comparative analysis reinforces the value of the proposed method as a significant improvement over previous approaches. It enhances classification accuracy and sensitivity and contributes to the development of explainable AI models in medicine. These advancements are critical for the future of precision medicine, where accurate, reliable, and interpretable diagnostic tools are essential for delivering the best possible

patient outcomes. The potential of the proposed method to revolutionise clinical practice is substantial, and its successful implementation could mark a major step forward in integrating AI into healthcare.

6.1 Summary of Key Findings

RQ1: Can an efficient deep learning colorectal polyp classification model be developed for insufficient training data using advanced data augmentation techniques?

Key Findings:

The study demonstrated that cDCGAN, unified with a concordant classifier, can enhance the dataset in terms of size, quality and diversity if the data available for training is insufficient. The results show that the average classification accuracy increased by 18% and sensitivity was enhanced by 17%, which is a substantial improvement over traditional data augmentation techniques such as flipping, rotation, cropping, etc, that improved the robustness and generalizability of the model.

RQ2: How can deep learning models be made more interpretable and transparent for CRC classification, ensuring that pathologists can trust and understand the decision-making process?

Key Findings:

Incorporating the Grad-CAM explainability method into the workflow allowed for visual explanations of the decisions taken. The regions of the images contributing towards the final decision were highlighted by the model, building the trust of pathologists in the developed system. A major factor that contributed positively

towards implementing an interpretable model is the extraction of fine detail-grain feature details which was achieved through the vision transformer.

RQ3: Can nonlinear dimensionality techniques regularise the CRC classification model and achieve a successful classification task?

Key Findings:

This study signifies that incorporating nonlinear dimensionality techniques in the pipeline, such as manifold learning can not only tackle the curse of dimensionality and handle the notorious overfitting problem, but it also plays a vital role in the regularisation of the model. Results obtained by implementing several manifold learning algorithms such as Isomap, LLE, Spectral Embedding, etc, indicate that the model was able to reduce the loss to a great extent, leading towards precise classification.

RQ4: Does the CNN classification model performance further enhance by developing custom-built networks for accurate colorectal polyp classification?

Key Findings:

A custom CNN network designed as the final classifier tailored to the network requirement further enhanced the classification performance as compared to off-the-shelf CNN models. In addition to improved performance, the complexity of the model was also reduced. Domain knowledge and constraints were directly incorporated into the custom network which led to reliability.

6.2 Implications of Research

The work presented in this study contributes to the growing body of literature on applying DNN to medical image analysis, focused on CRC polyp classification, providing evidence that by handling all the major challenges posed by biomedical images using DNN models, classification accuracy and sensitivity can be significantly enhanced. Furthermore, this research work offers insights into the significance of the interpretability of the model in clinical workflows.

The superior performance, as evidenced by increased accuracy, validates the effectiveness of the proposed method in handling the intricacies of CRC images. The findings suggest that implementing such models in clinical applications can lead to the early classification of polyps, providing better patient outcomes, and is also resourceful for clinicians in terms of time and effort. This virtual biopsy tool will be beneficial in a real clinical environment to decide on the severity of a patient's colorectal lesions when the patient is suffering from multiple lesions. Implementing a concordant synthetic image generation module, generalising the model, and incorporating explainable AI in the framework provided a transparent decision-making process.

6.3 Limitations and Future Work

Even though the developed framework provides promising results and the proposed methodology addresses several challenges posed by CRC data, potentially useful methods and directions were not explored due to the limited timeframe of our work. This work can be extended to diverse application domains where several future directions can be pinned down.

- **Scalability Issues:**

  Scaling this approach to larger datasets or high-resolution medical images presents several challenges. ViTs, due to their global self-attention mechanism, require exponentially increasing memory and computation as image resolution increases. Similarly, CGANs become harder to train on large datasets due to mode collapse risks and increased instability. Manifold learning techniques, especially those relying on pairwise distance calculations (e.g., t-SNE, Isomap), struggle with scalability as dataset size grows, leading to excessive computation time and memory usage. Strategies such as model pruning, mixed-precision training, and distributed computing are necessary to address these issues and enable large-scale deployment.

- **Training Time:**

  The combined use of ViTs, CGANs, and manifold learning results in significantly longer training times compared to conventional deep learning models such as CNNs. While CNNs benefit from efficient weight-sharing and local receptive fields, ViTs require extensive training with large-scale datasets to learn meaningful representations. CGANs, in turn, suffer from slow convergence due to the adversarial nature of training, which often requires careful hyperparameter tuning and extensive iterations to stabilise. Manifold learning techniques add another layer of complexity, especially when applied to high-dimensional medical data, as they require iterative computations to map data onto a lower-dimensional space while preserving structural integrity. Given these constraints, optimising training through GPU parallelisation, reduced-precision arithmetic (e.g., FP16 training), and transfer learning can help mitigate excessive computational demands and

improve training efficiency. Furthermore, transformer architectures are slow at training and prediction, require a large parameter count, and have a slow reasoning speed that becomes an issue in real-time tasks and large-scale rapid applications. Therefore, redesigning the model without impacting its performance is a challenging direction. This also makes the real-time detection of polyps or streaming data difficult. Therefore, designing such an effective method is one of the directions for this research.

- **High Computational Cost:**

  The proposed framework integrates Vision Transformers (ViTs), Conditional Generative Adversarial Networks (CGANs), and manifold learning, all of which are computationally intensive. ViTs demand substantial memory and processing power due to their self-attention mechanisms, which compute relationships between all image patches, leading to quadratic complexity with respect to input size. CGANs introduce additional computational overhead, as they involve training two competing networks: a generator synthesising realistic medical images and a discriminator distinguishing real from synthetic data, resulting in prolonged training cycles. Additionally, manifold learning, which is often used for dimensionality reduction or feature extraction, requires expensive distance computations to preserve the intrinsic structure of high-dimensional data. These factors collectively contribute to high computational costs, necessitating high-performance GPUs or TPUs to achieve feasible training times. In terms of the experiment environment, high-performance GPUs and TPUs such as NVIDIA Tesla are essential for large-scale datasets. It can provide a robust environment for large-scale data to classify images while maintaining the model's performance and efficiency.

Another challenge in the clinical setting is the confrontation with inaccurate data; hence, examining how to handle such data is the next direction of our research. One of the important directions is the utilisation of transformers and manifold learning in combination with the kNN graph and Reciprocal Graph to facilitate semi-supervised learning to improve classification where labelled data is scarce.

An important issue that was not addressed in this work is the potential risk posed by the enhanced capability of the synthetic image generation module of being misused to generate malicious content. This requires the development of more anti-forgery methods in the future to mitigate the risk of misuse and harness the benefit of synthetic image generation for positive use at the same time.

We have not focused on the classification of streaming data, the colonoscopy video generated in real time. Such data arrive continuously, and it is important that the model adapts to the requirements in real time and maintains performance. Moreover, in addition to polyps classification, it would be highly beneficial to incorporate the detection module into the framework before the classification module.

6.4 Road Map for Future Work

Addressing the limitations of the proposed framework and advancing its capabilities requires a strategic and collaborative approach. By implementing targeted improvements, we can improve its computational efficiency, interpretability, and generalisation, making it a more effective tool for colorectal cancer diagnosis and treatment. Ultimately, these enhancements will support clinical adoption, improve diagnostic accuracy, and contribute to better patient outcomes.

- **Improving Computational Efficiency:**

  The framework's reliance on computationally intensive models, such as Vision Transformers (ViTs) and Conditional Generative Adversarial Networks (CGANs), necessitates optimisation strategies to reduce resource consumption while maintaining performance. One effective approach is distributed training across multiple GPUs or TPUs, which enables parallel computation, significantly reducing training time and allowing the model to handle large-scale, high-resolution endoscopic images more efficiently. Additionally, techniques such as mixed-precision training (FP16/FP32), gradient checkpointing, and model pruning can further optimise memory usage and speed up training without compromising accuracy.

- **Enhancing Interpretability**

  For the framework to be clinically useful, its decision-making process must be transparent and interpretable for healthcare professionals. This requires a collaborative effort with clinicians to design user-friendly interfaces that present model outputs intuitively and meaningfully. Techniques such as saliency maps, Class Activation Mapping (CAM), and SHAP (Shapley Additive Explanations) can be integrated to highlight critical regions of medical images, helping doctors understand why a model makes a specific diagnosis. In addition, interactive visualisation tools that allow clinicians to explore the AI's reasoning process will foster trust and facilitate real-world deployment.

- **Contrast Enhancement Techniques**

  A common challenge in colorectal polyp detection is the misclassification of polyps due to texture similarities with normal mucosa, leading to false negatives. Contrast enhancement techniques can be incorporated into the pre-processing pipeline

to mitigate this issue and improve polyp visibility. Methods such as adaptive histogram equalisation (AHE), contrast-limited adaptive histogram equalisation (CLAHE), and deep learning-based enhancement models can help distinguish polyps more effectively. Furthermore, leveraging GAN-based image enhancement can generate high-quality, contrast-improved images to aid model training and clinical diagnosis.

References

Abbasi, A. A., Hussain, L. & Ahmed, B. (2022). Improving multi-class brain tumor detection using vision transformer as feature extractor. In *International Conference on Intelligent Systems and Machine Learning* (pp. 3–14).

Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E. & Alshammari, T. (2023). Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. *Big Data and Cognitive Computing*, *7*(2), 74.

Bello, I., Zoph, B., Vaswani, A., Shlens, J. & Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3286–3295).

Bengio, Y., LeCun, Y. et al. (2007). Scaling learning algorithms towards AI. *Large-scale Kernel Machines*, *34*(5), 1–41.

Bernal, J., Tajkbaksh, N., Sanchez, F. J., Matuszewski, B. J., Chen, H., Yu, L., ... others (2017). Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, *36*(6), 1231–1249.

Bjorck, N., Gomes, C. P., Selman, B. & Weinberger, K. Q. (2018). Understanding batch normalization. *Advances in Neural Information Processing Systems*, *31*.

Borhani, Y., Khoramdel, J. & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, *12*(1), 11554.

Brownlee, J. (2018). Ensemble learning methods for deep learning neural networks. *Machine Learning Mastery*, *18*.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M. & Zhou, C. (2020). Deep learning in medical image analysis. *Deep Learning in Medical Image Analysis: Challenges and Applications*, 3–21.

Chen, S.-A., Li, C.-L. & Lin, H.-T. (2021). A unified view of cgans with and without classifiers. *Advances in Neural Information Processing Systems*, *34*, 27566–27579.

Chicco, D. & Jurman, G. (2023). A statistical comparison between matthews correlation coefficient (MCC), prevalence threshold, and fowlkes–mallows index. *Journal of Biomedical Informatics*, *144*, 104426.

Corley, D. A., Jensen, C. D., Marks, A. R., Zhao, W. K., Lee, J. K., Doubeni, C. A., ... others (2014). Adenoma detection rate and risk of colorectal cancer and death. *New England Journal of Medicine*, *370*(14), 1298–1306.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Dolezal, J. M., Wolk, R., Hieromnimon, H. M., Howard, F. M., Srisuwananukorn, A., Karpeyev, D., ... others (2023). Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precision Oncology*, *7*(1), 49.

Dornaika, F. (2020). Multi-layer manifold learning with feature selection. *Applied Intelligence*, *50*(6), 1859–1871.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16 ×16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.

Dutta, P., Sathi, K. A., Hossain, M. A. & Dewan, M. A. A. (2023). Conv-ViT: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. *Journal of Imaging*, *9*(7), 140.

Elsheikh, T. (2008). Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *Am J Clin Pathol*, *130*, 683–686.

Geboes, K., Geboes, K. & Jouret-Mourin, A. (2013). Endoscopy and histopathology. *Endoscopy*, *1*, 3–32.

Gomes, H. M., Barddal, J. P., Enembreck, F. & Bifet, A. (2017). A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)*, *50*(2), 1–36.

Gschwantler, M., Kriwanek, S., Langner, E., Göritzer, B., Schrutka-Kölbl, C., Brownstone, E., ... Weiss, W. (2002). High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics. *European Journal of Gastroenterology & Hepatology*, *14*(2), 183–188.

Gu, Z., Li, Y., Wang, Z., Kan, J., Shu, J., Wang, Q. et al. (2023). Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention. *Computational Intelligence and Neuroscience*, *2023*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... others (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110.

Hee, K. N., Jung, Y. S., Jeong, W. S., Yang, H.-J., Park, S.-K., Choi, K. & Park, D. I. (2017). Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal Research*, *15*(3), 411–418.

Heinz, C. N., Echle, A., Foersch, S., Bychkov, A. & Kather, J. N. (2022). The future of artificial intelligence in digital pathology–results of a survey across stakeholder groups. *Histopathology*, *80*(7), 1121–1127.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Hsu, C.-M., Hsu, C.-C., Hsu, Z.-M., Shih, F.-Y., Chang, M.-L. & Chen, T.-H. (2021). Colorectal polyp image detection and classification through grayscale images and deep learning. *Sensors*, *21*(18), 5995.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Ishaq, S., Siau, K., Harrison, E., Tontini, G. E., Hoffman, A., Gross, S., ... Neumann, H. (2017). Technological advances for improving adenoma detection rates: The changing face of colonoscopy. *Digestive and Liver Disease*, *49*(7), 721–727.

Jae, Y., Bae, J. P., Kim, Chung, J.-W., Park, D. K., Kim, K. G. & Kim, Y. J. (2021). New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images. *Scientific Reports*, *11*(1), 3605.

Jha, D., Ali, S., Hicks, S., Thambawita, V., Borgli, H., Smedsrud, P. H., ... others (2021). A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging. *Medical Image Analysis*, *70*, 102007.

Jiseob, K., Shim, K., Kim, J. & Shim, B. (2023). Vision transformer-based feature extraction for generalized zero-shot learning. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5).

Kaminski, M. F., Regula, J., Kraszewska, E., Polkowski, M., Wojciechowska, U., Didkowska, J., ... Butruk, E. (2010). Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine*, *362*(19), 1795–1803.

Kaminski, M. F., Thomas-Gibson, S., Bugajski, M., Bretthauer, M., Rees, C. J., Dekker, E., ... others (2017). Performance measures for lower gastrointestinal endoscopy: A european society of gastrointestinal endoscopy (esge) quality improvement initiative. *United European Gastroenterology Journal*, *5*(3), 309–334.

Khazaee Fadafen, M. & Rezaee, K. (2023). Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework. *Scientific Reports*, *13*(1), 8823.

Kim, J., Hong, J. & Park, H. (2018). Prospects of deep learning for medical imaging. *Precision and Future Medicine*, *2*(2), 37–52.

Komeda, Y., Handa, H., Watanabe, T., Nomura, T., Kitahashi, M., Sakurai, T., ... others (2017). Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: Preliminary experience. *Oncology*, *93*(Suppl. 1), 30–34.

Kopelman, Y., Gal, O., Jacob, H., Siersema, P., Cohen, A. et al. (2019). Automated polyp detection system in colonoscopy using deep learning and image processing techniques. *Journal of Gastroenterology and its Complications*, *3*(1), 101.

Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., ... Hassanpour, S. (2017). Deep learning for classification of colorectal polyps on whole-slide images. *Journal of Pathology Informatics*, *8*(1), 30.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Kudo, S.-e., Mori, Y., Misawa, M., Takeda, K., Kudo, T., Itoh, H., ... Mori, K. (2019). Artificial intelligence and colonoscopy: Current status and future perspectives. *Digestive Endoscopy*, *31*(4), 363–371.

Kvålseth, T. O. (2017). On normalized mutual information: measure derivations and properties. *Entropy*, *19*(11), 631.

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., ... others (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. *Gastroenterology*, *134*(5), 1570–1595.

Li, K., Fathan, M. I., Patel, K., Zhang, T., Zhong, C., Bansal, A., ... Wang, G. (2021). Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *Plos One*, *16*(8), e0255809.

Liew, W. S., Tang, T. B., Lin, C.-H. & Lu, C.-K. (2021). Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches. *Computer Methods and Programs in Biomedicine*, *206*, 106114.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Lv, S., Shi, S., Wang, H. & Li, F. (2021). Semi-supervised multi-label feature selection with adaptive structure learning and manifold learning. *Knowledge-based Systems*, *214*, 106757.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L. & Bartoli, A. (2016). Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, *35*(9), 2051–2063.

Min, J. K., Kwak, M. S. & Cha, J. M. (2019). Overview of deep learning in gastrointestinal endoscopy. *Gut and Liver*, *13*(4), 388.

Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv:1411.1784*.

Misawa, M., Kudo, S.-e., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., ... others (2018). Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience. *Gastroenterology*, *154*(8), 2027–2029.

Mukhtorov, D., Rakhmonova, M., Muksimova, S. & Cho, Y.-I. (2023). Endoscopic image classification based on explainable deep learning. *Sensors*, *23*(6), 3176.

Nazir, S., Dickson, D. M. & Akram, M. U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, 106668.

Nogueira-Rodríguez, A., Domínguez-Carbajales, R., López-Fernández, H., Iglesias, Á., Cubiella, J., Fdez-Riverola, F., ... Glez-Pena, D. (2021). Deep neural networks approaches for detecting and classifying colorectal polyps. *Neurocomputing*, *423*, 721–734.

Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S. & Tada, T. (2020). Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therapeutic Advances in Gastroenterology*, *13*, 1756284820910659.

Pacal, I. & Karaboga, D. (2021). A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*, *134*, 104519.

Pacal, I., Karaboga, D., Basturk, A., Akay, B. & Nalbantoglu, U. (2020). A comprehensive review of deep learning in colon cancer. *Computers in Biology and Medicine*, *126*, 104003.

Park, M., Lee, M. & Yu, S. (2022). Hrgan: A generative adversarial network producing higher-resolution images than training sets. *Sensors*, *22*(4), 1435.

Pascotti Valem, L., Guimarães Pedronette, D. C. & Latecki, L. J. (2023). Graph convolutional networks based on manifold learning for semi-supervised image classification. , arXiv–2304.

Patino-Barrientos, S., Sierra-Sosa, D., Garcia-Zapirain, B., Castillo-Olea, C. & Elmaghraby, A. (2020). Kudo's classification for colon polyps assessment using a deep learning approach. *Applied Sciences*, *10*(2), 501.

Poudel, S., Kim, Y. J., Vo, D. M. & Lee, S.-W. (2020). Colorectal disease classification using efficiently scaled dilation in convolutional neural network. *IEEE Access*, *8*, 99227–99238.

Radford, A., Metz, L. & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, *34*, 12116–12128.

Rahman, M. M., Wadud, M. A. H. & Hasan, M. M. (2021). Computerized classification of gastrointestinal polyps using stacking ensemble of convolutional neural network. *Informatics in Medicine Unlocked*, *24*, 100603.

Raju, M. & Rao, B. S. (2022). Colorectal cancer disease classification and segmentation using a novel deep learning approach. *International Journal of Intelligent Engineering & Systems*, *15*(4).

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A. & Shlens, J. (2019). Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, *32*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rangarajan, A. K., Purushothaman, R. & Ramesh, A. (2018). Tomato crop disease classification using pre-trained deep learning algorithm. *Procedia Computer Science*, *133*, 1040–1047.

Rau, A., Edwards, P. E., Ahmad, O. F., Riordan, P., Janatka, M., Lovat, L. B. & Stoyanov, D. (2019). Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. *International Journal of Computer Assisted Radiology and Surgery*, *14*, 1167–1176.

Ren, Y., Tsai, M.-Y., Chen, L., Wang, J., Li, S., Liu, Y., ... Shen, C. (2020). A manifold learning regularization approach to enhance 3D CT image-based lung nodule classification. *International Journal of Computer Assisted Radiology and Surgery*, *15*, 287–295.

Rui, H. & Wu, Z. (2021). Multi-label feature selection via manifold regularization and dependence maximization. *Pattern Recognition*, *120*, 108149.

Sabol, P., Sinčák, P., Hartono, P., Kočan, P., Benetinová, Z., Blichárová, A., ... Jašková, A. (2020). Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of Biomedical Informatics*, *109*, 103523.

Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., ... Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, *46*(1), e1–e36.

Sakr, A. S., Soliman, N. F., Al-Gaashani, M. S., Pławiak, P., Ateya, A. A. & Hammad, M. (2022). An efficient deep learning approach for colon cancer detection. *Applied Sciences*, *12*(17), 8450.

Samad, M. D. & Sekmen, A. (2021). Manifold regularized convolutional neural network. In *International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1–6).

Sanchez-Peralta, L. F., Bote-Curiel, L., Picon, A., Sanchez-Margallo, F. M. & Pagador, J. B. (2020). Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artificial Intelligence in Medicine*, *108*, 101923.

Sasmal, P., Sharma, V., Bhuyan, M., Prakash, A. J., Patro, K. K., Samee, N. A., ... others (2023). Semi-supervised generative adversarial networks for improved colorectal polyp classification using histopathological images. *Information Sciences*, 120033.

Sener, O. & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. *arXiv:1708.00489*.

Shaban, M. T., Baur, C., Navab, N. & Albarqouni, S. (2019). Staingan: Stain style transfer for digital histological images. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (pp. 953–956).

Shen, D., Wu, G. & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248.

Shin, Y., Qadir, H. A. & Balasingham, I. (2018). Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, *6*, 56007–56017.

Siegel, R. L., Miller, K. D., Fuchs, H. E., Jemal, A. et al. (2021). Cancer statistics. *Ca Cancer J Clin*, *71*(1), 7–33.

Situ, J. (2023). Unsupervised image classifier based on manifold learning. *Transactions on Computer Science and Intelligent Systems Research*, *2*, 82–95.

Sobur, A. & Rana, I. C. (2024). Advancing cancer classification with hybrid deep learning: Image analysis for lung and colon cancer detection.

Sohail, A., Khan, A., Nisar, H., Tabassum, S. & Zameer, A. (2021). Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. *Medical Image Analysis*, *72*, 102121.

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P. & Vaswani, A. (2021). Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16519–16529).

Suzuki, K. (2012). A review of computer-aided diagnosis in thoracic and colonic imaging. *Quantitative Imaging in Medicine and Surgery*, *2*(3), 163.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B. & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Tan, J. W. & Jeong, W.-K. (2023). Histopathology image classification using deep manifold contrastive learning. *arXiv:2306.14459*.

Tandon, S. & Namboodiri, A. (2022). Transformer based fingerprint feature extraction. In *26th International Conference on Pattern Recognition (ICPR)* (pp. 870–876).

Tang, C., Bian, M., Liu, X., Li, M., Zhou, H., Wang, P. & Yin, H. (2019). Unsupervised feature selection via latent representation learning and manifold regularization. *Neural Networks*, *117*, 163–178.

Thany, S. (2010). Advances in experimental medicine and biology. *Springer: New York, NY, USA*, *683*.

Tharwat, M., Sakr, N. A., El-Sappagh, S., Soliman, H., Kwak, K.-S. & Elmogy, M. (2022). Colon cancer diagnosis based on machine learning and deep learning: modalities and analysis techniques. *Sensors*, *22*(23), 9250.

Trockman, A. & Kolter, J. Z. (2022). Patches are all you need? *arXiv:2201.09792*.

Tummala, S., Kadry, S., Nadeem, A., Rauf, H. T. & Gul, N. (2023). An explainable classification method based on complex scaling in histopathology images for lung and colon cancer. *Diagnostics*, *13*(9), 1594.

Ullman, S. (2000). *High-level Vision: Object Recognition and Visual Cognition*. MIT Press.

Vanitha, K., Sree, S. S. & Guluwadi, S. (2024). Deep learning ensemble approach with explainable ai for lung and colon cancer classification using advanced hyperparameter tuning. *BMC Medical Informatics and Decision Making*, *24*(1), 222.

Vanyan, A. & Khachatrian, H. (2021). Deep semi-supervised image classification algorithms: A survey. *JUCS: Journal of Universal Computer Science*, *27*(12).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vinh, N. X., Epps, J. & Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1073–1080).

Wagner, S. & Wagner, D. (2007). Comparing clusterings: An overview.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Wang, Z., Zhou, Y., Takagi, T., Song, J., Tian, Y.-S. & Shibuya, T. (2023). Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data. *BMC Bioinformatics*, *24*(1), 139.

Wei, J. W., Suriawinata, A. A., Vaickus, L. J., Ren, B., Liu, X., Lisovsky, M., ... others (2020). Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. *JAMA Network Open*, *3*(4), e203398–e203398.

WEI, Q. (2023). *Computational methods for deep learning: Theory, algorithms, and implementations*. SPRINGER VERLAG, SINGAPOR.

Xia, L., Mi, S., Zhang, J., Luo, J., Shen, Z. & Cheng, Y. (2023). Dual-stream feature extraction network based on cnn and transformer for building extraction. *Remote Sensing*, *15*(10), 2689.

Xuejun, Zhang, S., Bu, Z., Ma, L. & Huang, J. (2023). Texture feature dimensionality reduction-based mammography classification using random forest. *Journal of Computational Methods in Sciences and Engineering*(Preprint), 1–9.

Yan, W. Q. (2019). *Introduction to intelligent surveillance: surveillance data capture, transmission, and analytics*. Springer.

Yang, L., Yu, X., Huang, J. X. & An, A. (2011). Combining integrated sampling with svm ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617–631.

Yangxi, L., Shi, X., Du, C., Liu, Y. & Wen, Y. (2016). Manifold regularized multi-view feature selection for social image annotation. *Neurocomputing*, *204*, 135–141.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal ofPhysics: Conference Series* (Vol. 1168, p. 022022).

Yongzhen, H., Wu, Z., Wang, L. & Tan, T. (2013). Feature coding in image classification: A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 493–506.

Yoon, D., Kong, H.-J., Kim, B. S., Cho, W. S., Lee, J. C., Cho, M., ... others (2022). Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network. *Scientific Reports*, *12*(1), 261.

Younas, F., Usman, M. & Yan, W. Q. (2023). A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*, *53*(2), 2410–2433.

Yu, H. & Zhang, X. (2020). Synthesis of prostate MR images for classification using capsule network-based GAN model. *Sensors*, *20*(20), 5736.

Zachariah, R., Samarasena, J., Luba, D., Duh, E., Dao, T., Requa, J., ... Karnes, W. (2020). Prediction of polyp pathology using convolutional neural networks achieves "resect and discard" thresholds. *The American Journal of Gastroenterology*, *115*(1), 138.

Zafar, I., Tzanidou, G., Burton, R., Patel, N. & Araujo, L. (2018). *Hands-on convolutional neural networks with tensorflow: Solve computer vision problems with modeling in tensorflow and python*. Packt Publishing Ltd.

Zhang, Ruikai, Zheng, Y., Mak, T. W. C., Yu, R., Wong, S. H., ... Poon, C. C. (2016). Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 41–47.

Zhang, J., Luo, Z., Li, C., Zhou, C. & Li, S. (2019). Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, *95*, 136–150.

Zhang, Y., Kang, B., Hooi, B., Yan, S. & Feng, J. (2023). Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10795–10816.

Zhang, Y., Ma, Y. & Yang, X. (2022). Multi-label feature selection based on logistic regression and manifold learning. *Applied Intelligence*, 1–18.

Zhao, Hengshuang, Jia, J. & Koltun, V. (2020). Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10076–10085).

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y. & Han, S. (2020). Differentiable augmentation for data-efficient GAN training. *Advances in Neural Information Processing Systems*, *33*, 7559–7570.