MEDICAL IMAGE ANALYSIS OF WIRELESS CAPSULE ENDOSCOPY DATA

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

in

Electronics and Communication Engineering

by

PALAK HANDA

(Enrollment No.: 2K21/PHDEC/05)

Under the supervision of

PROF. INDU SREEDEVI
Professor in ECE Deptt., and Dean (Student Welfare), DTU

PROF. NIDHI GOEL
Professor in ECE Deptt., IGDTUW

To the
Department of Electronics and Communication
Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
July, 2024

© DELHI TECHNOLOGICAL UNIVERSITY, DELHI, 2024

ALL RIGHTS RESERVED

ACKNOWLEDGEMENTS

"Great work requires great and persistent effort for a long time……Character has to be established through a thousand stumbles."

Swami Vivekananda

Reflecting upon the past three years, I find so many people who have been instrumental in the completion of this research work. It is a matter of great privilege for me to acknowledge the help, guidance and encouragement which I have received from several quarters during the research period.

First and foremost, thanks to the Almighty for giving me strength and inspiration to carry out this research work. I owe a deep sense of gratitude to all his comprehensive soul whose divine light has enlightened my path throughout the journey of my research.

I take the opportunity to humbly submit my sincere and heartfelt thanks to my gurus (research supervisors), **Prof. S. Indu** and **Prof. Nidhi Goel** from the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, and the Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi for their invaluable guidance, enthusiastic encouragement, and persistent support. I am truly grateful from the core of my heart for their meticulous approach, wonderful assistance of their perspective, and fruitful discussions on my research topic. Their immense contribution and rare dedication in providing the much-needed guidance, is worth of highest honor. Their careful supervision and personal attention have given me a lot of confidence and enthusiasm during the different stages of my doctoral investigations. They are academic giants under whose watch am molded to a seasoned research scholar. I invariably fall short of words to express my sincere gratitude for their patience and motivation.

I am immensely thankful to **Dr. Deepak Gunjan** from the Department of Gastroenterology and HNU, All India Institute of Medical Sciences New Delhi for his invaluable guidance, support and ground-breaking medical expertise throughout this journey. Sir’s encouragement and contrastive feedback have shaped this research and my understanding in the field of gastroenterology and its potential applications using artificial intelligence.

I extend my sincere appreciation to the members of my doctoral committee for their insightful comments and thoughtful suggestions that have enhanced the quality of this thesis. Thank you, **Prof. Dinesh K. Vishwakarma**, Head of the Department of Information Technology for the motivation, support and insightful comments and suggestions throughout this dissertation and my master’s program. I am extremely thankful to **Prof. O. P. Verma**, Head of the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, and other faculty members for their endless support and cooperation throughout this dissertation. I am thankful to all staff members, especially **Mr. Manish**, of the department of Electronics and Communication Engineering for their kind help and support during the entire period of my research. I am grateful to the management of Delhi Technological University, Delhi, Indira Gandhi Delhi Technical University for Women, Delhi, and All India Institute of Medical Sciences New Delhi for providing the necessary resources, facilities, and a conducive research environment for my research work. The virtual platform Medical Imaging and Signal Analysis (MISAHUB) has been a source of inspiration and collaboration for my research work. I am thankful to all the mentees (+100) and medical experts (+8) at MISAHUB.

I express my sincere gratitude towards **Prof. Pravat K. Mandal** from the Neurospectroscopy and Neuroimaging Laboratory, National Brain Research Centre Gurugram for his invaluable guidance and mentorship provided on how to conduct clinical trials and perform statistical analysis. The hands-on experience and real-world examples shared by sir have been exceptionally beneficial in bridging the gap between biomedical theory and its application in healthcare. I am also thankful to **Dr. Sandeep Jaiswal** from the Department of Biomedical Engineering, Mody University of Science and Technology, Rajasthan for growing the seed of research inside me and inspiring me to do better in life. Thank you, **Dr. Chundawat**, I have learnt so much from your human biology sessions.

I am greatly indebted to my dear friends **Ms. Apurva, Mrs. Aditi, Mr. Pulin, Mr. Ajay, Ms. Mehar, Ms. Harsha, Ms. Priyanka, Ms. Dimple**, and **Ms. Swati** who constantly supported, counselled, and encouraged me all the time throughout this research period. Thank you, Mr. Hari Om for all the firsts. Those important life lessons will always stay with me. Thank you, Late Pulkit and Sir Paul Richard Alexander for inspiring me to stay positive and follow the light in life.

My appreciation also goes to my colleagues in the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, and Indira Gandhi Delhi Technical University for Women, Delhi for their constant laughter, lunches, support and assistance especially **Dr. Rajiv, Mrs. Divya, Mrs. Aapurva, Mrs. Shivani, Mr. Vijay, Ms. Tanvika, Mrs. Barkha, Ms. Kavita, Mrs. Monika, Mrs. Muskan, Mr. Vishal, Ms. Kirti, Mr. Arjun, Mr. Anil, Ms. Himanshi, Mr. Nihal, Mr. Rahul, Mrs. Ramsha, Mrs. Vidhi, Ms. Pallavi, Dr. Monika**, and **Dr. Ruchika**.

I dedicate this thesis to my family and teachers for their endless love, support, encouragement, and blessings throughout my 26 years of existence. My parents, **Mr. Parveen Handa, Mrs. Geeta Handa**, my sister, **Ms. Chaitanya Handa** and my god mother **Ms. Nirmal Kanta** without whom I could not imagine being enrolled in Ph.D., for their unwavering encouragement in the moments of despair and discouragement, undeterred faith in me since childhood, and being the biggest pillar of strength who supported me all the way till the end. I feel very lucky to have a great family including my maternal and paternal grandparents, uncles, aunts, and cousins who have supported me throughout the journey of my research work. Thank you, kiwi and coco, for being there.

Last but not least, I sincerely acknowledge each one of those who directly or indirectly have or have not helped, rejected, jibed, and overlooked me during the whole period, thus making it a well-rounded, strong experience of learning. This thesis is a culmination of the collective efforts of all those who have touched my academic and personal life, and for that, I am truly grateful.

Palak Handa

PALAK HANDA

DELHI TECHNOLOGICAL UNIVERSITY
Formerly Delhi College of Engineering
Shahbad Daulatpur, Main Bawana Road, Delhi –42

CANDIDATE'S DECLARATION

I **Palak Handa** hereby certify that the work which is being presented in the thesis entitled **Medical Image Analysis of Wireless Capsule Endoscopy Data** in partial fulfillment of the requirements for the award of the Degree of Doctor in Philosophy, submitted in the **Department of Electronics and Communication Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from **August 2021** to **June 2024** under the supervision of **Prof. S. Indu** and **Prof. Nidhi Goel**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Palak Handa

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY
Formerly Delhi College of Engineering
Shahbad Daulatpur, Main Bawana Road, Delhi –42

CERTIFICATE BY THE SUPERVISOR(S)

Certified that **Palak Handa** (Enrollment No.: 2K21/PHDEC/05) has carried out their research work presented in this thesis entitled "**Medical Image Analysis of Wireless Capsule Endoscopy Data**", for the award of **Doctor of Philosophy** from the Department of Electronics and Communication Engineering, Delhi Technological University, under our guidance and supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Prof. Indu Sreedevi
Supervisor
Department of ECE
Delhi Technological University,
Delhi –110042, India

Prof. Nidhi Goel
Co-Supervisor
Department of ECE
Indira Gandhi Delhi Technical University for Women,
Delhi –110006, India

Sumantra Dutta Roy
Indian Examiner

This thesis is dedicated to my parents,
Mr. Parveen Handa, Mrs. Geeta Handa,
my sister, Ms. Chaitanya Handa, and
my god mother, Ms. Nirmal Kanta.

For their endless love,
support and encouragement

Shri Krishna says,

"Bas karm tumhara kal hoga. Aur karm mein agar sachai hai, to karm kaha nisfal hoga.

Har ek sankat ka hal hoga. Vo aaj nahi to kal hoga.

Loha jitna tapta hai, utna hi taakat bharta hai. Sone ko jitni aag lag, vo utna prakhar nikharta hai.

Heere par jitni dhaar pade, vo utna khoob chamakta hai. Mitti ka bartan pakta hai, tab dhun par khoob khanakta hai.

Sooraj jaisa banna hai to sooraj jitna jalna hoga. Nadiyo sa aadar paana hai to parvat chodh nikalna hoga.

Aur hum aadam ke bete hai, kyu soche raah saral hoga. Kuch zyada vakt lagega par sangharsh zaroor safal hoga.

Har ek sankat ka hal hoga. Vo aaj nahi to kal hoga."

~ Radhey Govinda

MEDICAL IMAGE ANALYSIS OF WIRELESS CAPSULE ENDOSCOPY DATA

PALAK HANDA

ABSTRACT

Gastrointestinal (GI) diseases, often diagnosed through endoscopy, constitute a significant global health burden. The manual inspection of endoscopy data, particularly in colonoscopies and video capsule endoscopy (VCE) is time-intensive and prone to oversight. Automating abnormality detection and cleanliness assessment through medical image analysis (MIA) and artificial intelligence (AI) promises to revolutionize this process, offering quicker and more precise diagnostics. Such assessments may help to enhance patient outcomes by developing sophisticated algorithms capable of detecting abnormalities and assessing cleanliness in real-time. By streamlining endoscopy evaluations, such automatic assessments may help in addressing critical healthcare needs, facilitating earlier detection, and intervention for GI diseases, ultimately improving patient care and reducing the burden on healthcare systems.

Currently, the diagnostic yield of colonoscopy and VCE in a real-time clinical setting has been investigated in several Indian and abroad medical studies but MIA and AI based studies to perform automatic abnormality detection and cleanliness assessment in endoscopy are rare. The absence of high quality, multi-labelled, and medically validated AI datasets is a major reason behind the less no. of studies being conducted. Lack of AI datasets hinder a transparent comparison between the performance of existing automated systems in this field with the up-coming systems. Most of the automated systems have been designed for less no. of endoscopy frames which are unavailable for public research use. The existing datasets mostly contain binary class labels such as ‘abnormal’ or ‘normal/healthy’ and ‘clean’ or ‘unclean/dirty’ and ‘adequate’ or ‘inadequate’ and do not provide information related to mucosal visual quality, presence of impairments, artefacts, medical scores and distortions etc.

Multi-label classification is an emerging and presently less explored area. It

has the potential to address several tasks such as the automatic cleanliness scoring in endoscopy. The assessment of cleanliness in endoscopy is crucial for ensuring optimal visualization and, consequently, accurate diagnosis. Cleanliness metrics play a pivotal role in maintaining the quality of endoscopy examinations, allowing healthcare professionals to make informed decisions based on clear and unobstructed images. They play an even more crucial role in VCE as it is non-invasive in nature and lacks therapeutic capabilities. Owing to the above-discussed research gaps, this research focuses on two tasks namely automatic detection of abnormality in polyp and non-polyp frame in colonoscopy frames and automatic assessment of cleanliness in VCE.

The first task focused on developing an explainable, end-to-end and robust architecture for automatic colorectal polyp diagnosis using colonoscopy polyp and non-polyp frames. The developed architecture consisted of a novel, fine-tuned feature-extracting module, followed by polyp and non-polyp frame identification and a window-based polyp detection system. To show the robustness of the developed architecture, a new test set was developed and evaluated. After the analysis, it was released on Zenodo, an open-source platform for research purposes. It is called the gastrointestinal atlas-colon polyp dataset. It consisted of seven patient videos obtained from open-source, copyright free web sources. Explainable and evaluation methods like class activation mapping, feature mapping, occlusion testing, hyper-parameter tuning ablation experiments, and separate, sequential, and non-sequential frame-based test set analysis have been used to show the efficacy of the proposed architecture. The developed architecture has been compared with the existing state-of-the-art methodologies in this field. Additionally, architecture has been compared with a transfer learning architecture as well.

The second task focused on development of methodology to automatically assess the cleanliness in VCE video frames. First, the process of scoring VCE frames has been automated as per existing KOrea CanaDA (KODA) scoring system. The process is an easy-to-use mobile application called AI-KODA. AI-KODA Score is a flutter-based application which can be downloaded on a mobile. The application first trains a gastroenterologist how to use KODA. After a simple training, the

gastroenterologist can upload VCE video frames on the application and score them. After successful scoring, a report is generated for the overall score. The scores are also collected in real-time and saved for the development of a frame level, high-quality, and multi-labelled dataset for automatic multi-label classification of clean v/s dirty VCE video frames. The developed dataset has been subjective to medical verification with the help of three experienced gastroenterologists. Based on the common consensus by the three gastroenterologists, a common dataset comprising of 2173 with eight distinct labels of KODA has been developed. A comprehensive evaluation, interpretation, benchmarking of the generated dataset has been done using ten machine learning models and eight transfer learning algorithms on google Collaboratory and a supercomputer named, NVIDIA RTX A5000 workstation. The developed dataset and its methodology are first-of-its-kind.

The proposed methodologies for these two tasks are scalable, robust, work in real-time, and are explainable in nature. The comprehensive analysis followed for each of the tasks shows its promising future in the gastroenterology department. Both the methodologies help in reducing the time and effort of the gastroenterologist in timely detection of polyps in colonoscopy and cleanliness assessment in VCE.

LIST OF PUBLICATIONS

Journal Papers

- P. Handa *et al.*, “Automatic Detection of Colorectal Polyps with Mixed Convolutions and Its Occlusion Testing”, *Neural Comput Appl*, vol. 35, no. 26, pp. 19409–19426, 2023. (accepted and published)
- P. Handa *et al.*, “Comprehensive Evaluation of a New Automatic Scoring System for Cleanliness Assessment in Video Capsule Endoscopy”, *Int J Imaging Syst Technol*, vol. 34, no. 3, p. e23097, 2024. (accepted and published)
- P. Handa *et al.*, “A Multi-Label Dataset and Its Evaluation for Automated Scoring System For Cleanliness Assessment In Video Capsule Endoscopy”, *Phys Eng Sci Med*, pp. 1–14, 2024. (accepted and published)
- P. Handa *et al.*, “AI-KODA Score Application for Cleanliness Assessment in Video Capsule Endoscopy Frames”, *Minimally Invasive Therapy & Allied Technologies.* (accepted and in-production)

Conference Papers

- P. Handa, N. Goel, and S. Indu, “Datasets of Wireless Capsule Endoscopy For AI-Enabled Techniques”, in *International Conference on Computer Vision and Image Processing*, Springer, 2021, pp. 439–446 (accepted and published)
- P. Handa, N. Goel, and S. Indu, “Automatic Intestinal Content Classification Using Transfer Learning Architectures”, in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, 2022, pp. 1–5 (accepted and published)

Book Chapters

- P. Handa *et al.*, “Computer-aided polyp detection using customized Convolutional Neural Network architecture,” *in Intelligent Data Analytics for Bioinformatics and Biomedical Systems*, Wiley. (accepted and in-production)

Released Datasets

- P. Handa *et al.*, “Test data of published work titled 'Automatic Detection of Colorectal Polyps with Mixed Convolutions and its Occlusion Testing'”, Neural Computing and Applications, vol. 35. Zenodo, pp. 19409–19426, Jun. 27, 2023. doi: 10.1007/s00521-023-08762-z. (open-sourced)
- P. Handa *et al.*, “AI-KODA Dataset: An AI-Image Dataset for Automatic Assessment of Cleanliness in Video Capsule Endoscopy as per KOrea-CanaDA Scores”. figshare. Dataset. May 13, 2024. doi:10.6084/m9.figshare.25807915.v1. (open-sourced)

TABLE OF CONTENTS

Acknowledgment iii-v
Candidate's Declaration vi
Certificate by the Supervisor(s) vii
Abstract x-xii
List of Publications xiii-xiv
Table of Contents xv-xvii
List of Tables xviii-xix
List of Figures xx-xxiii
List of Abbreviations xxiv-xxvi

Chapter 1: Introduction 1-13

1.1 Endoscopy
  1.1.1 Colonoscopy
  1.1.2 Video Capsule Endoscopy
1.2 Challenges in Assessment of Endoscopy Data
1.3 Role of Artificial Intelligence in Assessment of Endoscopy Data
1.4 Research Motivation
1.5 Problem Formulation
1.6 Research Objectives
1.7 Research Contribution
1.8 Outlines of the Thesis

Chapter 2: Literature Review 14-36

2.1 Introduction
2.2 Endoscopy Datasets Available for Artificial Intelligence-enabled Techniques
2.3 Abnormality Detection Methodologies in Endoscopy
2.4 Cleanliness Assessment Methodologies in Endoscopy
2.5 Multi-label Classification in Endoscopy
2.6 Chapter Summery and Gaps in the Study

2.7 Conclusion and Future Scope

Chapter 3: Development of Deep Learning Architecture for Abnormality Detection in Endoscopy 37-58

3.1 Introduction
3.2 Related Works
3.3 Problem Statement
3.4 Methodology
  3.4.1 Data preparation and augmentation
  3.4.2 Proposed Architecture
  3.4.3 Experimental Settings
3.5 Results and Discussion
  3.5.1 Hyper-parameter tuning
  3.5.2 Feature mapping, class activation mapping and ablation experiments
  3.5.3 Test set analysis and occlusion testing
  3.5.4 Comparison with vanilla Inception v3 architecture
  3.5.5 Comparison with existing state-of-the-art works
3.6 Conclusion and Future Scope

Chapter 4: Development of Artificial Intelligence Korea Canada for Cleanliness Assessment in Endoscopy 59-72

4.1 Introduction
4.2 Related Works
4.3 Problem Statement
4.4 Methodology
  4.4.1 Preparation of capsule videos
  4.4.2 Study design
  4.4.3 Statistical Analysis Setup
4.5 Results
  4.5.1 Demographic and statistical details of the patients and the developed dataset

4.5.2 AI-KODA Score Application Feedback and Efficacy
4.5.3 Reliability estimates
4.6 Discussion
4.7 Conclusion and Future Scope

Chapter 5: Development of Classification Techniques for Cleanliness Assessment in Endoscopy 73-99

5.1 Introduction
5.2 Related Works
5.3 Problem Formulation
5.4 Methodology
5.5 Results and Discussion
5.6 Conclusion and Future Scope

Chapter 6: Conclusion, Future Scope and Social Impact 100-102

6.1 Summary
6.2 Future Scope/ Directions and Social Impact

References 103-117

List of Publications 118-121
Plagiarism Report 122-123
Curriculum Vitae/ Brief Profile 124-127

LIST OF TABLES

Table No.	Title	Page No.
Table 2.1	List of open-source colonoscopy polyp datasets with available links.	15
Table 2.2	Comparative analysis of the available colonoscopy datasets in this field.	22
Table 2.3	Comparative analysis between the existing, available VCE datasets for research use.	25
Table 3.1	Details of the chosen dataset.	40
Table 3.2	Reported results of the accuracy and loss for various varied hyper-parameters in the proposed feature extracting module.	48
Table 3.3	Results of the test set analysis.	53
Table 3.4	Comparison of the proposed pipeline with existing state-of-the-art works. NR = not reported.	57
Table 4.1	Average measure of inter-rater intra-class correlation coefficients for five-minute interval and random intervals in twenty-eight capsule videos.	68
Table 4.2	Average measure of intra-rater intra-class correlation coefficients for five-minute interval and random intervals in twenty-eight capsule videos.	69
Table 5.1	Previous approaches in this field.	76
Table 5.2	Details of the transfer learning algorithms used in the three multi-label classification tasks.	87
Table 5.3	Achieved evaluation metrics on validation data.	88
Table 5.4	Achieved evaluation metrics on test data.	89
Table 5.5	Achieved evaluation metrics on training and validation data for only VM labels.	90-91

Table 5. 6	Achieved evaluation metrics on training and validation data for only OV labels.	92-93
Table 5.7	Achieved evaluation metrics on training and validation data for both OV and VM labels.	93-94
Table 5.8	Achieved evaluation metrics on test data for only OV, only VM, and both OV and VM labels.	94-95

LIST OF FIGURES

Figure No.	Title	Page No.
Figure 1.1	Graphical representation of the google web searches of the keyword ‘Gastrointestinal disease’ from 2004-2023 across globe. The numbers on the y axis represent search interest relative to the highest point on the chart for the given region and time. X axis represents the time period considered.	1
Figure 1.2	Incidence and mortality of GI cancer cases in 2018 [6].	2
Figure 1.3	Representative diagram of colonoscopy procedure. Source: Bio render developed.	3
Figure 1.4	Representative diagram of a video generated from a capsule endoscopy. The video contains abnormal findings (polyp) in stomach region. Source: Video segment collected from Deptt. Of Gastroenterology and HNU, AIIMS Delhi.	4
Figure 1.5	Graphical representation of the google web searches of the keyword ‘capsule endoscopy’, ‘colonoscopy’, and ‘AI in healthcare’ from 2004-2022 across globe. The numbers on the y axis represent search interest relative to the highest point on the chart for the given region and time. X axis represents the time period considered. Source: Google trends.	6
Figure 1.6	Representative diagram to depict the research motivation behind this thesis.	9
Figure 1.7	Representative diagram to depict the problem formulation approach utilized in this thesis.	9
Figure 3.1	Examples of the colorectal polyp and non-polyp frames considered for training, validating, and testing the proposed architecture. Train polyp data (a) Kvasir v1, (b)	42

Etis-Larib, (c) CVC-Colon; Validation polyp data (d) Kvasir v1, (e) Etis-Larib, (f) CVC-Colon; Testing data (g) Proposed Gastrointestinal atlas-Colon Polyp test data, (h) Mixture (i) Polyp-Gen data, (j) Developed frames for occlusion testing. The non-polyp frames in training and validation data are from Kvasir v1 as Etis-Larib and CVC-Colon don't contain colorectal non-polyp frames.

Figure 3.2 Reported results of the accuracy and loss for various varied hyper-parameters in the proposed feature extracting module. 47

Figure 3.3 Achieved feature maps of the different convolution layers. (a) Depthwise separable layer, (b) Depthwise layer, (c) Transpose layer, and (d) Conventional convolution layer. 49

Figure 3.4 Achieved feature maps of the proposed feature extracting module. 50

Figure 3.5 Class activation maps. 51

Figure 3.6 Missed polyp frames during test set analysis from Polyp-gen dataset.

Figure 3.7 Colorectal polyp (CP) frame examples of the results achieved during test set analysis. (a) Correctly detected CP frames. (b) Repeatedly missed CP frames. (c) and (d) Correctly detected and identified frames. (e) Occlusive frames. 54

Figure 3.8 Achieved loss and AUC score graphs for the proposed architecture and vanilla Inception v3 architecture. 55

Figure 4.1 Example of CE frames selected over (a) – (h) five-minute interval and (i) – (p) random interval in one capsule video using the RAPID viewer software. 63

Figure 4.2 Snapshot of the of the developed application. (a) Login page, (b) home page, (c) training module, and (d) testing module with its different operations. 65

Figure 4.3 Frontend and Backend Data Flow Storage System of AI-KODA Score Application. 65

Figure 4.4 KODA score reference sheet. 66

Figure 4.5 Demographics of the twenty-eight capsule videos. (a) Types of diseases diagnosed, (b) Types of anomalies found. 68

Figure 4.6 (a) – (f) Sample frames with ambiguous scores. In case of (a), the opinion varied between <5% and 5%-25% for percentage of obstructed view. In case of (b), the opinion varied between 5%-25% and 26%-50% for percentage of obstructed view. In (c), the opinion varied between <5% and 5-25% for percentage of obstructed view. Some ambiguities were found for the percentage of mucosa visualized. In (d) the opinion varied between >75% and 25%-49% for the percentage of mucosa visualized. Ambiguous scores of percentages of mucosa visualized and obstructed view were achieved in both (e) and (f). For (f), the opinion varied between 50%-75% and 25%-49% for mucosa visualized and 5%-25% and >50% for obstructed view. Similarly, in (f) the opinion varied between <25% and 50%-75% for mucosa visualized and >50% and <5% for obstructed view. 71

Figure 5.1 Label distribution of the developed dataset. 79

Figure 5.2 Correlation matrix of the labels. 80

Figure 5.3 Gradient activation map of a VCE frame with 25% visualized mucosa and 50% obstructed view for the three multi-label classification pipelines. 96

Figure 5.4 Gradient activation map of a VCE frame with 75% visualized mucosa and 5% obstructed view for the three multi-label classification pipelines. 96

LIST OF ABBREVIATIONS

GI	Gastrointestinal
WHO	World Health Organization
CAGR	Compound Annual Growth Rate
VCE	Video Capsule Endoscopy
MIA	Medical Image Analysis
AI	Artificial Intelligence
KODA	KOrea CanaDA
GLRC	Gastrointestinal lesions in regular colonoscopy
KUMC	University of Kansas Medical Centre
SVM	Support Vector Machine
ROC	Region Over Curve
RF	Random Forest
KNN	K Nearest Neighbours
LR	Logistic Regression
MICCAI	Medical Image Computing and Computer Assisted Intervention
SSD	Single Shot Detector
CNN	Convolutional Neural Networks
GPNet	Global Pooling Network
CAD	Computer-Aided Detection
FP	False Positive

YOLO	You Look Only Once
FDA	Food and Drug Administration
RGB	Red-Green-Blue
CAC	Computer-aided Assessment of Cleansing
GLCM	Grey Level Correlation Matrix
Inception ResNet V2	ner residual network version 2
AUC	Area Under Curve
LSTM	Long Short-Term Memory
WADT-MCPI	Window bAsed Detection afTer Mixed Convolutions Polyp Identification
VGG	Visual Geometric Group
ReLu	Rectified linear activation function
CP	Colorectal polyps
SIFT	Scale-invariant feature transform
HOG	Histogram of Oriented Gradients
DNN	Deep Neural Networks
1-D	One-Dimensional
2-D	Two-Dimensional
3-D	Three-Dimensional
Grad-CAM	Gradient activated class activation map
U net	U network
SegNet	Segment network
SGD	Stochastic Gradient Descent

SHAP	Shapley Additive exPlanations
TP	, True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
ICCs	Intra-class Correlation Coefficients
CI	Confidence Interval
R/G	Red over Green pixel
MLP	Multi-layer Perceptron
DT	Decision Tree
SVC	Support Vector Classifier
NB	Naïve Bayes
VM	Visualized Mucosa
OV	Obstructed View
MobileNet	Mobile Network
DenseNet	Dense Network

CHAPTER 1

INTRODUCTION

The gastrointestinal (GI) tract, or digestive tract or alimentary canal, is the tract or passageway of the digestive system which extends from the mouth to the anus [1]. The GI tract contains all the major organs of the digestive system, in humans and other animals, including the mouth, pharynx (throat), oesophagus, stomach, small intestine, large intestine, rectum, and anus [1]. GI diseases found in these organs are widespread across the globe. Various factors brought on by industrialization, changes in nutrition and diet, and increased use of antibiotics have contributed in increasing the prevalence rate of GI diseases over the years [2], [3], [4]. Figure 1.1 represents the growing popularity of the keyword 'gastrointestinal disease' on google web searches over the last 20 years worldwide.

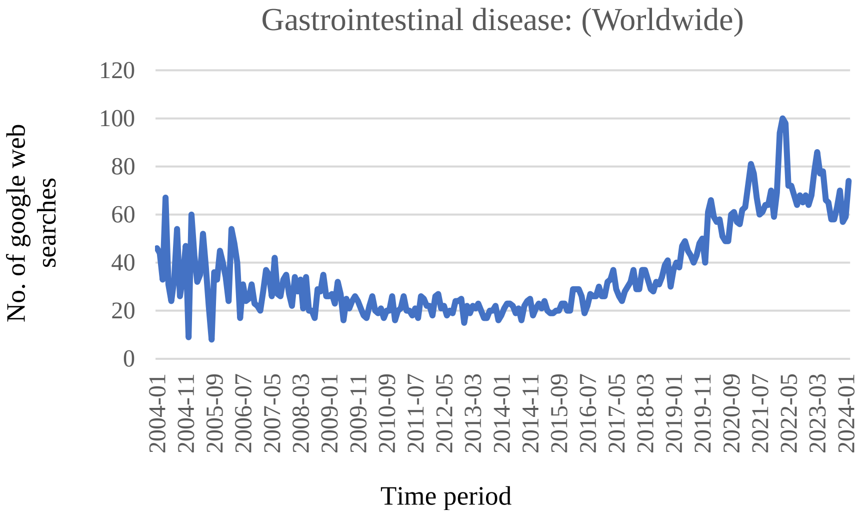


Figure 1.1 Graphical representation of the google web searches of the keyword 'Gastrointestinal disease' from 2004-2023 across globe. The numbers on the y axis represent search interest relative to the highest point on the chart for the given region and time. X axis represents the time period considered. Source: Google trends.

According to the latest statistics provided by World Health Organization (WHO), about 40% of the population across globe suffer from GI diseases at some point of their lives [5]. In the United States, digestive diseases affect more than 40 million individuals and account for millions of clinical visits annually, with health care expenditures totalling \$119.6 billion in 2018 [5]. Cancers of the GI have grown and become responsible for one in four of the found cancer cases and account for one in three cancer deaths globally (Figure 1.2) [2], [3], [5], [6].

Different type of endoscopy methods is being utilized to diagnose, treat and manage these diseases. Due to the growing prevalence of GI diseases across globe, the global market of endoscopy has exponentially increased. In the past five years, the Compound Annual Growth Rate (CAGR) of global endoscopy visualization systems like colonoscopy and video capsule endoscopy (VCE) have been up by 7.2% and account to more than a 100 billion market revenue in the next ten years [7], [8].

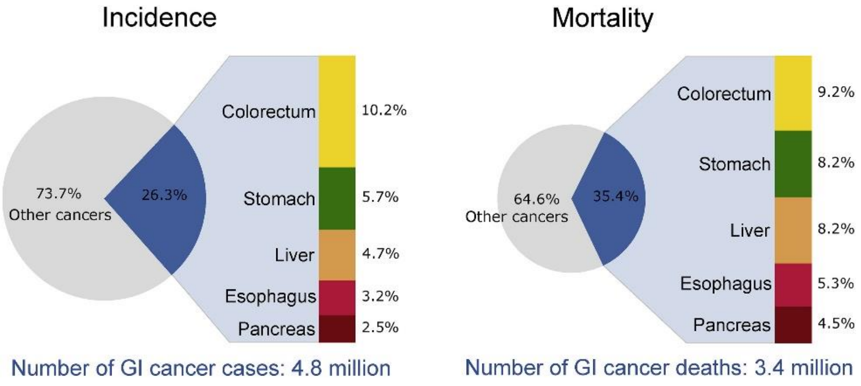


Figure 1.2 Incidence and mortality of GI cancer cases in 2018 [6].

1.1 Endoscopy

Endoscopy is a medical procedure that allows a doctor to inspect and observe the inside of the body, commonly the GI tract. It is of various types depending upon the type of organs considered for examination. The varied types of endoscopies allow for real-time examination, aiding in the identification of abnormalities, obtaining biopsies, and facilitating therapeutic interventions for diagnosis and treatment purposes of GI diseases. In this thesis, we will focus on primarily two types of endoscopies namely colonoscopy and VCE.

1.1.1 Colonoscopy

Colonoscopy is a diagnostic and therapeutic endoscopy method to examine the large intestine, and the distal portion of the small intestine with the help of a colonoscope

(Figure 1.3) [9]. A colonoscope is a flexible, slender tube equipped with a light source and camera. It is a crucial diagnostic tool for detecting colorectal conditions, such as polyps, tumours, or inflammation [9], [10]. It remains a gold standard for colon cancer screening which accounts for major cancer in GI [11], [12], [13]. Medtronic, Ethicon (Johnson and Johnson), Boston Scientific, Olympus, R. Bard, Coloplast, Hologic, Bayer, Applied Medical, Cook Medical are some of the famous colonoscopy manufacturing companies.

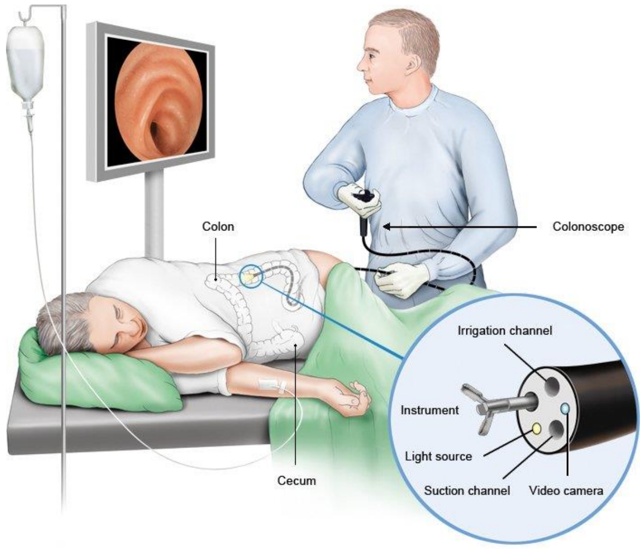


Figure 1.3 Representative diagram of colonoscopy procedure. Source: Bio render developed.

1.1.2 Video Capsule Endoscopy

VCE is relatively a newer, non-invasive endoscopic method which allows direct visualization of the GI tract especially small bowel due to its miniaturized size, and safeguarding sedation-related complications in diseased individuals [14]. It consists of a disposable capsule-shaped device which comprises of an optical dome, a battery, an illuminator, an imaging sensor, and a radio-frequency transmitter [14]. Figure 1.4 depicts the VCE of stomach region. VCE is often considered as an alternative to conventional endoscopy methods [15]. Therapeutic Endoscopy Guidelines and Recommendations 2018-19 released by Indian Association of Gastrointestinal Endo Surgeons have recommended the use of VCE as an initial test for stable patients with overt or occult small-bowel bleeding and Crohn’s disease [16]. CapsoVision, Checkcap, Chongqing Jinshan Science and Technology, and Given Imaging are some of the famous companies of VCE.

1.2 Challenges in Assessment of Endoscopy Data

Both colonoscopy and VCE have improved a physician’s ability to find different abnormalities in the GI tract such as Crohn’s disease, Celiac disease, worms, polyps, and ulcer, etc [17], [18], [19]. However, these endoscopy methods are yet to reach its true potential in developing countries like India due to its cost, lack of standardized procedure for cleanliness assessment, poor reproducibility, and longer reading time without compromising the quality of the report with high false positives [10], [15], [17], [20], [21], [22].

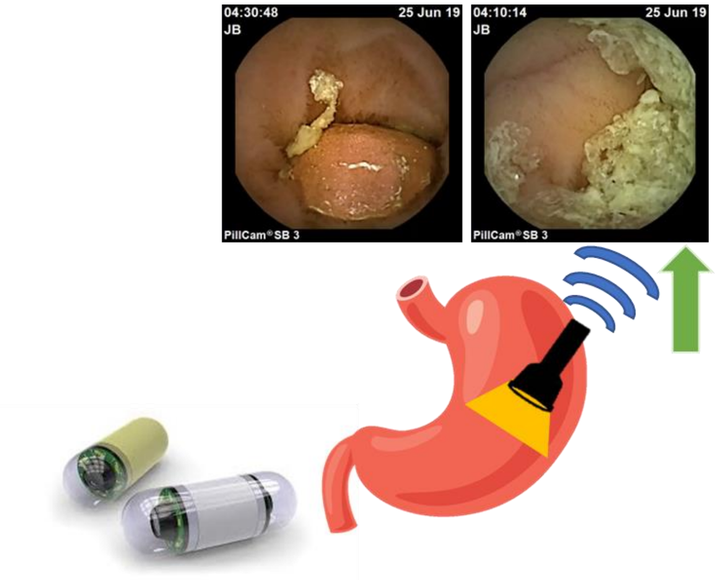


Figure 1.4 Representative diagram of a video generated from a capsule endoscopy. The video contains abnormal findings (polyp) in stomach region. Source: Video segment collected from Deptt. Of Gastroenterology and HNU, AIIMS Delhi.

In terms of average cost, colonoscopy and VCE are approximately Rs. 3,000 - 10,000 and Rs. 25,000 - 40,000 respectively which is difficult to afford by common men in developing counties like India. In terms of cleanliness assessment, Boston scoring is preferred for colonoscopy. Presently there is no standardized protocol followed before VCE to assess its cleanliness across the globe.

Approximately 2−3 hours of reading time is taken by an experienced gastroenterologist to inspect these endoscopy videos of 2-12 hours through a frame-by-frame analysis. Manual interpretation and chances of the false-positive rate of the abnormalities in the endoscopy videos are high due to impairment of the mucosal frames with bubbles, debris, intestinal fluid, foreign objects, and chyme (food), etc [10], [15], [17], [20], [21], [22]. Qualification of the gastroenterologists, their experience in performing the endoscopy methods, measure of fatigue, distraction, and ability to multi-task in busy clinical schedules also contributes to higher miss-rate of the abnormalities. Further, poor bowel preparation, adequate mucosal exposure, contrast, and lightening and other hardware related technological limitations delays this process [10], [17], [23].

1.3 Role of Artificial Intelligence in Assessment of Endoscopy Data

There has been an active participation from medical practitioners, researchers and industry professionals to perform medical image analysis (MIA) and artificial intelligence (AI) for the automatic assessment of endoscopy data [18], [24], [25], [26], [27]. MIA and AI are predicted to have profound effects on the future of colonoscopy and VCE in the context of automatic abnormality detection, segmentation, classification, cleanliness assessment, scoring system, depth estimation, odometry, video summarization, and artefact removal systems [24], [28], [29]. They are believed to be free from fatigue, distraction, and other human biases. MIA focuses on development of methods and process of imaging the interior of a body for visual representations, clinical decision-making analysis, research, and medical intervention. Along with the integration of AI, such automated assessments can aid in reducing the burden on gastroenterologists and save their valuable time by reducing the inspection time of endoscopy video analysis while maintaining high diagnostic precision and improved reproducibility. Figure 1.5 depicts the popularity of the keywords like ‘capsule endoscopy’, ‘colonoscopy’, and ‘AI in healthcare’ over the last 20 years worldwide.

In AI tasks, a classification task refers to a type of labelling where an image or video is assigned certain concepts. The AI model tries to classify input data and predict new data based on the assigned concepts (class labels). Based on the class label information in a dataset, it can be a binary, multi-class, multi-label or hierarchal classification. In the context of endoscopy, a binary classification may help the researcher to identify whether a video frame is ‘abnormal’ or ‘normal’ or ‘clean’ or ‘unclean’. A multi-class classification may help in identifying the name of the abnormality like ‘polyp’, ‘ulcer’, ‘tumour’, ‘worms’, and ‘vascular malformation’, etc. A multi-label classification task may help in identifying the grade of cleanliness of a particular video frame like ‘excellent’, ‘good’, ‘fair’ and ‘poor’ which may be mutually exclusive of each other’s labels. Classification task may also be referred to as classifying a particular abnormality, for example polyp according to its medical morphology and types like Paris classification, and Kudo’s classification.

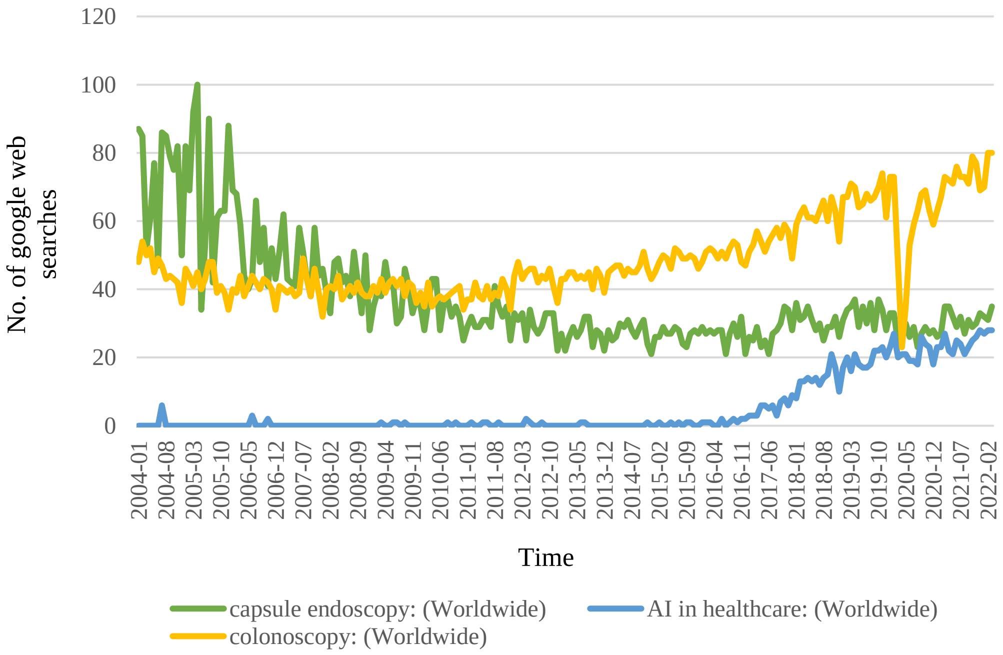


Figure 1.5 Graphical representation of the google web searches of the keyword ‘capsule endoscopy’, ‘colonoscopy’, and ‘AI in healthcare’ from 2004-2022 across globe. The numbers on the y axis represent search interest relative to the highest point on the chart for the given region and time. X axis represents the time considered. Source: Google trends.

Visual data quality assessment task includes assessing the cleanliness of the GI tract, its intestinal content classification and overall quality assessment in terms of resolution, specularity, pacemaker artefacts, saturation, blurring and contrast of an endoscopy video frame or segment. Video summarization task aims to generate a short summary of the content of a longer endoscopy video by selecting and presenting the most informative video segments to the gastroenterologists. Localization task aims to find the location of the video segment or frame in the GI tract. Abnormality detection task includes identifying an abnormality along with its location in the video frame with the help of bounding box and class label information. It may also help in distinguishing between various abnormalities, intestinal fluid, and mucosal layers, lumen etc., in a particular video segment or a frame.

Segmentation is a type of labelling where each pixel in an image is labelled with pre-defined concepts. An image may be divided into pixel groupings which may be then labelled and classified. This is done with the goal of simplifying an image or changing how an image is presented to the AI model, to make it easier to analyze and interpret. In the context of endoscopy frame, a segmentation task may aim to provide the exact outline of the abnormality and or grade of cleanliness within a video frame or segment. A pixel-by-pixel details may be provided for a given abnormality and or grade of cleanliness, as opposed to classification models, where the model identifies what is in an abnormality and or grade of cleanliness in an endoscopy video frame, and detection models, which places a bounding box around specific abnormalities and or grade of cleanliness. Segmentation tasks are computationally more expensive and require detailed medical information in comparison to the above list tasks.

1.4 Research Motivation

GI diseases, often diagnosed through endoscopy, constitute a significant global health burden [6]. The manual inspection of endoscopy data, particularly in colonoscopies and VCE, is time-intensive and prone to oversight [30]. Automating abnormality detection and cleanliness assessment through MIA and AI promises to revolutionize this process, offering quicker and more precise diagnostics [31], [32]. Such assessments may help to enhance patient outcomes by developing sophisticated algorithms capable of detecting abnormalities and assessing cleanliness in real-time. By streamlining endoscopy evaluations, such automatic assessments may help in addressing critical healthcare needs, facilitating earlier detection, and intervention for

GI diseases, ultimately improving patient care and reducing the burden on healthcare systems [30], [31], [32], [33], [34], [35].

Currently, the diagnostic yield of colonoscopy and VCE in a real-time clinical setting has been investigated in several Indian and abroad medical studies but MIA and AI based studies to perform automatic abnormality detection and cleanliness assessment in endoscopy are rare. The absence of high quality, multi-labelled, and medically validated AI datasets is a major reason behind the less no. of studies being conducted. Lack of AI datasets hinder a transparent comparison between the performance of existing automated systems in this field with the up-coming systems. Most of the automated systems have been designed for less no. of endoscopy frames which are unavailable for public research use. The existing datasets mostly contain binary class labels such as ‘abnormal’ or ‘normal/healthy’ and ‘clean’ or ‘unclean/dirty’ and ‘adequate’ or ‘inadequate’ and do not provide information related to mucosal visual quality, presence of impairments, artefacts, medical scores and distortions etc.

Multi-label classification is an emerging and presently less explored area. It has the potential to address several tasks such as the automatic cleanliness scoring in endoscopy. The assessment of cleanliness in endoscopy is crucial for ensuring optimal visualization and, consequently, accurate diagnosis. Cleanliness metrics play a pivotal role in maintaining the quality of endoscopy examinations, allowing healthcare professionals to make informed decisions based on clear and unobstructed images. They play even more crucial role in VCE as it is non-invasive in nature and lacks therapeutic capabilities.

With progressing research in AI and development of these automated systems, interpretability and extensive performance analysis also needs to be addressed. In this thesis, two tasks namely automatic detection of abnormality in polyp and non-polyp frame in colonoscopy frames and automatic assessment of cleanliness in VCE will be focused. Figure 1.6 depicts the representative diagram of the research motivation behind this thesis.

1.5 Problem Formulation

The field of endoscopy plays a crucial role in diagnosing and monitoring various GI conditions. However, the manual assessment of endoscopy data is time-consuming, subjective, and prone to human error [31], [32], [36]. The integration of MIA and AI has shown promise in automating this process, offering potential improvements in efficiency and accuracy. This thesis aims to develop and evaluate an AI-based system

.

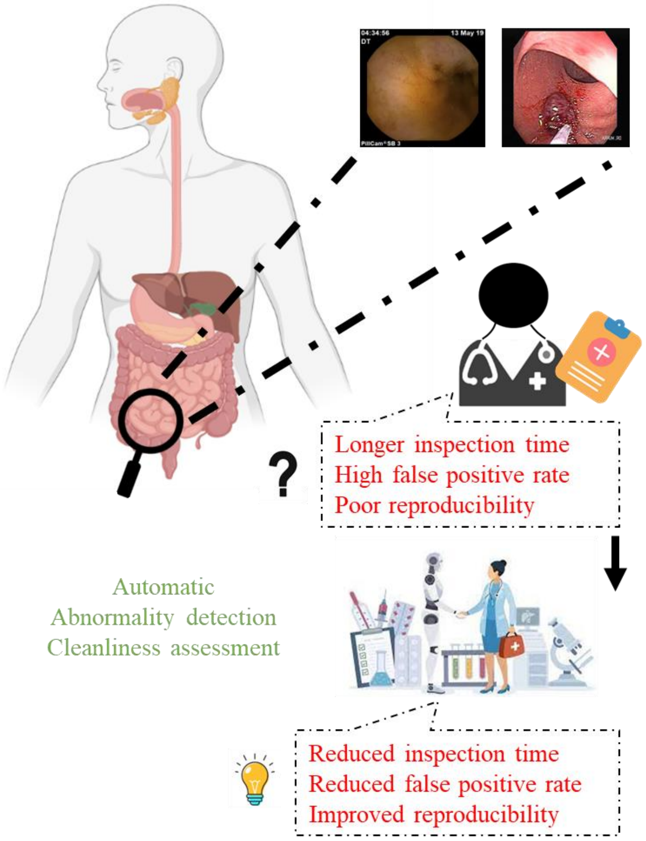


Figure 1.6 Representative diagram to depict the research motivation behind this thesis.

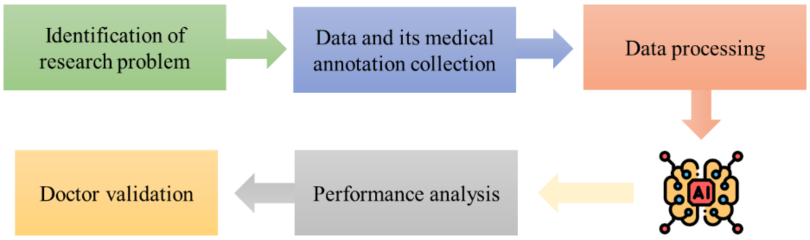

Identification of research problem
Data and its medical annotation collection
Data processing
AI
Performance analysis
Doctor validation

Figure 1.7 Representative diagram to depict the problem formulation approach utilized in this thesis.

for the automatic assessment of endoscopy data, primarily for colonoscopy and VCE, addressing key challenges in automatic abnormality detection and cleanliness assessment. Endoscopy data collection and its medical validation with the help of experienced gastroenterologists is vital in this field. It is followed by its data processing and development of AI-based system for automatic abnormality detection and cleanliness assessment. Performance analysis and result verification with the help of experienced gastroenterologists are very important. Figure 1.7 shows a representative diagram of the problem formulation for this thesis. In the subsequent chapters, several steps will be added in this.

1.6 Research Objectives

The primary goal of this thesis is to develop methodologies to perform automatic detection of abnormality in polyp and non-polyp frame in colonoscopy frames and automatic assessment of cleanliness in VCE. The methodologies used, its application, research findings, and accomplishments for each of the research goals are listed below in this segment:

Research Objective 1:

- To develop an endoscopy dataset for research use.

Research Objective 2:

- To develop a deep learning architecture for reducing the diagnostic time for abnormality detection in the developed endoscopy dataset.

Research Objective 3:

- To develop a transfer learning-based approach to classify and assess the cleanliness of the developed frames.

Research Objective 4:

- To develop a deep learning architecture for multi-label classification pipeline for the developed dataset.

1.7 Research Contribution

This thesis has focused on two tasks namely automatic detection of abnormality in polyp and non-polyp frame in colonoscopy frames and automatic assessment of cleanliness in VCE. The first task focused on developing an explainable, end-to-end and robust architecture for automatic colorectal polyp diagnosis using colonoscopy polyp and non-polyp frames. The developed architecture consisted of a novel, fine-tuned feature-extracting module, followed by polyp and non-polyp frame identification and a window-based polyp detection system. To show the robustness of the developed architecture, a new test set was developed and evaluated. After the analysis, it was released on Zenodo, an open-source platform for research purposes. It is called the gastrointestinal atlas-colon polyp dataset. It consisted of seven patient videos obtained from open-source, copyright free web sources. Explainable and evaluation methods like class activation mapping, feature mapping, occlusion testing, hyper-parameter tuning ablation experiments, and separate, sequential, and non-sequential frame-based test set analysis have been used to show the efficacy of the proposed architecture. The developed architecture has been compared with the existing state-of-the-art methodologies in this field. Additionally, the architecture has been compared with a transfer learning architecture as well.

The second task focused on development of methodology to automatically assess the cleanliness in VCE video frames. First, the process of scoring VCE frames has been automated as per existing KOrea CanaDA (KODA) scoring system. The process is an easy-to-use mobile application called AI-KODA. AI-KODA Score is a flutter-based application which can be downloaded on a mobile. The application first trains a gastroenterologist how to use KODA. After a simple training, the gastroenterologist can upload VCE video frames on the application and score them. After successful scoring, a report is generated for the overall score. The scores are also collected in real-time and saved for the development of a frame level, high-quality, and multi-labelled dataset for automatic multi-label classification of clean v/s dirty VCE video frames. The developed dataset has been subjective to medical verification with the help of three experienced gastroenterologists. Bases on the common consensus by the three gastroenterologists, a common dataset comprising of 2173 with eight distinct labels of KODA has been developed. A comprehensive evaluation, interpretation, benchmarking of the generated dataset has been done using ten machine learning models and eight transfer learning algorithms on google Collaboratory and a super computer named, NVIDIA RTX A5000 workstation. The developed dataset and its methodology are first-of-its-kind.

The proposed methodologies for these two tasks are scalable, robust, work in real-time, and are explainable in nature. The comprehensive analysis followed for each of the task, shows its promising future in the gastroenterology department. Both the methodologies help in reducing the time and effort of the gastroenterologist in timely detection of polyps in colonoscopy and cleanliness assessment in VCE.

1.8 Outlines of the Thesis

The thesis entitled, **'Medical Image Analysis of Wireless Capsule Endoscopy Data'** comprises six chapters followed by conclusions and future scope and a bibliography. The thesis is organized as following:

Chapter 1: Introduction

This chapter will cover the motivation and purpose of the outlined research topic. It will also contain the main idea for the development of the thesis.

Chapter 2: Literature Review

In this chapter, a detailed literature review of the available endoscopy datasets in this field for MIA and AI, existing methodologies for automatic detection of abnormality and cleanliness assessment in endoscopy primarily for colonoscopy and VCE will be done. Existing endoscopy studies utilizing multi-label classification tasks will also be discussed. It will be followed by discussion of the current limitations of the existing studies, chapter summary and contribution of the present research to the field.

Chapter 3: Development of Deep Learning Architecture for Abnormality Detection in Endoscopy

This chapter contributes to the development of a deep learning architecture abnormality detection in endoscopy. In this, polyp and non-polyp has been dealt as abnormalities. The chapter details the utilized methodology to prepare the AI dataset, its processing, and experimental settings to execute the proposed deep learning architecture for polyp and non-polyp detection in colonoscopy frames. It is followed by discussion of the achieved results, its comparison with the existing methodologies in this field and future scope in the field.

Chapter 4: Development of Artificial Intelligence Korea Canada for Cleanliness Assessment in Endoscopy

This chapter contributes to the development of an easy-to-use mobile based application called AI-KODA for real-time data collection and medical annotation for an automatic cleanliness assessment in endoscopy primarily VCE. The chapter details the utilized methodology to develop the mobile application, prepared dataset, method of annotation, study design and statistical analysis of the study. It is followed by discussion of the achieved results, its comparison with the existing methodologies in this field, efficacy of the developed application, reliability estimates, and conclusion and future scope in the field.

Chapter 5: Development of Classification Techniques for Cleanliness Assessment in Endoscopy

This chapter contributes to the development of classification techniques for automatic assessment of cleanliness in endoscopy primarily VCE. The chapter details the utilized methodology to prepare the AI dataset, its processing, and experimental settings to execute the classification tasks. It is followed by discussion of the achieved results, its comparison with the existing methodologies in this field and future scope in the field.

Chapter 6: Conclusion, Future Scope and Social Impact

This chapter will contain the summary of all the ideas, observations and contributions of the results obtained in each objective. Also, the future directions in this field are sketched in this section.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In response to challenges faced in the tiring analysis of endoscopy data, researchers and clinicians have turned to MIA and AI as a transformative solution [31], [33], [37]. This chapter details the literature review of this field. It will unfold in a structured manner, beginning with the discussion of the existing datasets available in this field. It will be followed by an exploration of the foundational studies that laid the groundwork for automated endoscopy analysis. It will then progress to recent developments in AI algorithms, deep learning techniques, and computer vision applications specifically tailored for abnormality detection and cleanliness assessment in colonoscopy and VCE.

The significance of this literature review lies in providing a comprehensive understanding of the state-of-the-art methodologies, and challenges associated with automated endoscopy analysis. By synthesizing existing knowledge, this chapter aims to guide future research endeavours, fostering the evolution of more robust, accurate, and clinically applicable AI systems for endoscopy. Ultimately, the integration of such technologies into routine clinical practice has the potential to revolutionize GI diagnostics, leading to earlier detection, improved patient outcomes, and more efficient healthcare delivery.

2.2 Endoscopy Datasets Available for Artificial Intelligence-enabled Techniques

Recent advancements in AI techniques are enabling automation in gastroenterology field especially for the detection, localization, segmentation and classification of colon polyps, associated lesions, abnormal growth and bleeding in the colon and rectum region which helps a physician to screen and diagnose colorectal cancer at an early stage [18], [24], [26], [38], [39]. High-quality, open access and free colonoscopy data plays an important role in escalating research and advancements in this field [27].

Various colonoscopy data with different medical settings, medically verified annotations, masks and bounding boxes have been released by researchers since 2012 [27], [29], [40]. We discuss each of the dataset one by one. Table 2.1 details the names of the dataset, their available link and the no. of polyp and non-polyp frames in the dataset.

Table 2.1 List of open-source colonoscopy polyp datasets with available links.

S. No.	Dataset	No. of polyps and Non-polyps	Application
1	Polypgen Database	3762 and 2520 (images)	Polyp detection and segmentation
2	Kvasir	1000 and 1000 (images)	Classification
3	Kvasir SEG	1000 and 0 (images)	Polyp segmentation
4	Hyper Kvasir	1028 (images)	Classification and segmentation
5	CVC Clinic	612 and 0 (images)	Polyp detection and segmentation
6	CVC Colon	380 and 0	Polyp detection
7	CVC EndosceneStill	912 and 0	Polyp segmentation
8	CVC PolypHD	56 and 0	Polyp segmentation
9	Etis-Larib	196 and 0	Polyp detection
10	ASU Mayo	5200 and 14200	Polyp recognition
11	CVC-ClinicVideoDB	38 (videos)	Polyp detection
12	Piccolo	3433 and 0	Polyp classification
13	CP-CHILD	1400 and 8100	Polyp detection
14	PIBAdb	31400 and 14000	Polyp classification
15	CVC300	300 and 0 (images)	Polyp detection
16	SUN Colonoscopy Dataset	49,136 and 109,554 (frames)	Colorectal-polyp detection
17	LDPolyp Video Database	200 and 0 (images)	Polyp detection
18	BKAI-IGH NeoPolyp Database	1000 and 0 (images)	Polyp segmentation and detection
19	Endotest	48641 and 205113 (frames)	Polyp recognition
20	GLRC Dataset	76 and 0 (videos)	Classification
21	Colonoscopic Dataset	155 and 0 (videos)	Polyp recognition and detection

CVC Colon dataset consists of 15 short sequences of colonoscopy videos from 13 patients which are annotated by medical experts [41]. The dataset contains 380 polyp images in total along with binary masks as ground truth. In each sequence exactly one polyp is shown and the frame size captured is 500×574 pixels. The dataset was developed as part of the GIANA grand challenge of 2017 and 2018. CVC Colon can be used for tasks like polyp segmentation and detection.

Etis-Larib database was developed by the Universitat Autonoma de Barcelona [42]. It contains 196 images from endoscopy videos including one polyp in each image. The data was collected from 44 different polyps in 34 sequences [42]. Images have a resolution of 1225×966 pixels. Annotations are also provided in from of binary masks (polyp locations) by medical specialists.

ASU-Mayo clinic database consists of a total of 19,400 frames with 5200 polyps and 14,200 without polyps [43]. 286 patients from 11 different centres were part of the study and data was collected from 10 positive and 10 negative shots of polyps, at multiple scales and camera angles [43]. Ground truth is provided in form of a binary mask (white for polyp region) which was created by expert colonsocopists. Colonoscopy images include narrow band imaging and vary greatly in their appearance, including the level of colon preparation, the colonoscopic events, and the amount of motion and interlacing artifacts.

Gastrointestinal lesions in regular colonoscopy (GLRC) dataset consists of 76 colonoscopy videos which were 30 seconds long and contain ground truth of histopathology, endoscopist inspection and camera calibration for 3D shape reconstruction [44]. The dataset contains 15 serrated adenomas, 21 hyperplastic lesions and 40 adenomas. The data was gathered using white light imaging and narrow band imaging, recording the lesion from various viewpoints and angles [44]. The aim of building the dataset was to help replicate experiments and compare various machine learning approaches for classification and detection along with help in automated classification of adenoma and serrated adenoma. The effect of shape features on classification can also be studied using the colonoscopic dataset by employing structure for motion techniques.

CVC ClinicDB dataset consists of 612 images from 31 colonoscopy sequences with a resolution of 384×288 [45]. The ground truth is provided in the form of a mask. The dataset has been utilized for medical image segmentation tasks related to polyp detection in colonoscopy videos. It was the official database used in the medical image computing and computer assisted intervention (MICCAI) 2015 Sub-Challenge on Automatic Polyp Detection Challenge in Colonoscopy Videos [46].

CVC-EndoScene still dataset is a combination of two datasets namely CVC- ColonDB and CVC-ClinicDB [47]. It consists of 44 video sequences (912 white-light images) which are extracted from 36 colonoscopy patient videos [47]. In addition to the images, the dataset also provides additional annotations (hand-made) for lumen and specular highlight segmentation along with defining a void class for the black boundaries of the images. Four classes in total are thus annotated namely polyp, lumen, background, specularity and border as void. The dataset is split into 60% training, 20% validation and 20% test sets with no patient overlapping in any set (20 patients in train, 8 validation and 8 in test). It has been released to provide a benchmark of endoluminal scene segmentation.

CVC-PolypHD provides 56 polyps images in high definition and is widely used in segmentation tasks [41], [47]. Binary masks of polyp locations are included as ground truths and the resolution of images is 1920×1080 pixels in comparison with the SD resolutions of 574×500 and 384×288 pixels. It was a part of the GIANA grand challenge polyp segmentation task (validation set) held in 2017 and 2018.

CVC-ClinicVideoDB was introduced as part of the GIANA grand challenge 2017 and 2018 for colonoscopy polyp detection [41]. The database includes 38 short and long sequence videos. 18 SD videos are provided for training with resolutions of 768×576 pixels. Binary masks for polyp locations are provided as ground truth. The data is collected from patients at Hospital Clinic of Barcelona, Spain.

CVC-300 is a 300-colonoscopy image database. It contains annotations of all sequences showing polyps with frame size of 500×574 pixels [48]. The data is collected from 13 colonoscopy patients. Different types of polyp appearance can be found in this database, leading to a great deal of variation. This database was part of the GIANA grand challenge polyp detection and segmentation task of 2017 and 2018 as a test set.

The Kvasir dataset (v1) contains 4000 images from inside the GI tract [49]. The images are annotated by experts and include 8 classes with anatomical landmarks like Z-line, pylorus, cecum, etc. and pathological findings including esophagitis, polyps, ulcerative colitis, etc. Additionally, several images related to polyp removal like ‘dyed and lifted polyp’, ‘dyed resection margins’, etc. are also provided. Images with resolutions of 720×576 and 1920x1072 pixels are present and the data is collected at the Vestre Viken Health Trust (VV) in Norway. By using an electromagnetic imaging system (ScopeGuide, Olympus Europe), some classes of images have a green box in the corner of the image to illustrate the endoscope’s position and configuration inside the bowel. The main applications of this dataset are automatic detection and classification of pathological findings in endoscopy procedures. Version 2 of the dataset, also called as Kvasir v2, consists of meticulously annotated and augmented images from the version 1. This dataset was part of the Mediaeval Medical Multimedia Challenge and made available in 2017. 33536 total images are contained in Kvasir v2 with 4192 images in each of the 8 classes. It is further divided into 80% train and 20% validation sets.

Hyper Kvasir contains about one million gastrointestinal tract image and videos [50]. The data is collected from gastric and colonoscopy examinations from Baerum Hospital, Norway between 2008 to 2016. 110,079 (10,662 labelled and 99,417 unlabelled) images and 374 videos are present with anatomical, normal and pathological findings [50]. Authors have provided a bounding box, and a segmentation mask, of 1000 polyp images. The motivation behind creating the database was to provide a large dataset for colonoscopy-based machine learning research as the existing databases are small. Moreover, providing partially labelled data can aid supervised as well as semi and unsupervised learning solutions.

The Kvasir SEG dataset contains annotated polyp images from the original Kvasir v1 dataset [51]. Authors have developed their corresponding masks. Some of the original images contain the image of the endoscope position marking probe from the

ScopeGuide (Olympus). Each folder contains 1000 images. The Kvasir SEG dataset has image folder, masks folder and JSON file (for bounding boxes). It is suitable for general segmentation, bounding box detection, localization, and classification of polyps. It can also assist the development of robust solutions for images captured by colonoscopies from different manufacturers.

Piccolo dataset comprises of 3433 images (2131 White-light intensity and 1302 narrow bind intensity) collected from Hospital Universitario Basurto (Bilbao, Spain) [52]. The images are extracted from 40 patients and 76 lesions [52]. The dataset is manually annotated with a binary mask for polyps along with a void class for not required background by medical experts. Clinical metadata is also provided in a CSV file format. 854×480 and 1920×1080 are the 2 image resolutions available [52]. Olympus endoscopes (CF-H190L and-CF-HQ190L) are used for video capturing and the data is divided into training (2203), validation (897) and test (333) sets with no set having same patients.

Polypgen database is a comprehensive polyp detection and segmentation database consisting of colonoscopy data from more than 300 patients admitted in 6 different data centres situated in Paris, Italy, Norway, UK and Egypt [53]. The dataset includes both single frames split as well as sequence data of positive (polyp containing) and negative (absence of polyp) samples [53]. A major limitation of the data is that the sequence positive samples contain a mixture of polyp instances and normal mucosa frames due to continuity of polyp appearance and disappearance. The negative samples also contain cases of colon linings, light reflections and mucosa covered with stool which can be mistaken for polyps. A total of 6282 frames are present in this dataset with 3762 positive samples and 2520 negative frame samples with bounding box annotations in VOC format [53].

CP-CHILD dataset collected at Hunan children’s hospital contains colonoscopy images of children under 18 years [54]. A total of 1600 children’s data is recorded and used to create two polyp datasets i.e., CP-CHILD-A and CP-CHILD-B. CHILD- A dataset contains 7000 non polyp and 1000 polyp images while CHILD-B includes 1100 non polyp and 400 polyp RGB images [54]. CHILD-A images are taken by Olympus

PCF-H290DI, while FUJIFLIM EC-530 wide meter is used for CHILD-B. Image resolution is 256×256 and class labels are identified by endoscopists. The training set contains 800 images of non-polyps and 300 images of polyps, while the test set contains 100 images of polyps and 300 images of non-polyps.

Polyp Image Bank database included colorectal polyp images and collected using white light as well as narrow band imaging. A histological report of each polyp is also provided. Annotations are done manually by clinical experts in the form of bounding boxes. A total of 31,400 polyp images (22,600 white light and 8,800 narrow band imaging) and 14,000 non-polyp images are present in the dataset from 1,176 different polyps. Video and image resolution is 768×576. This dataset can be used for polyp classification for different kinds of polyps like Adenoma vs. Hyperplastic vs. Sessile Serrated Adenoma vs. Traditional Serrated Adenoma vs. Non-Epithelial Neoplastic vs. Invasive.

SUN Colonoscopy Dataset is used to evaluate the effectiveness of an automated colorectal polyp detection system, based on colonoscopy videos [55]. It is collected at the Showa University Northern Yokohama Hospital. It contains still images from 113 colonoscopy videos, 100 being positive (containing polyp) and 13 negative samples [55]. There are 49,136 polyp frames in the SUN database, each one annotated with bounding boxes. 109,554 frames include non-polyp scenes. There are many different types of polyps in the resulting database i.e., 7 hyperplastic polyps, 4 sessile serrate lesions, 82 low-grade adenomas, 2 traditional serrated adenomas, 4 high-grade adenomas, and 1 invasive cancer.

LDPolypVideo Database contains video footage from colonoscopies showing polyps and more complex bowel environments is included in this large-scale database [56]. The database consists of 160 colonoscopy videos in which 40,266 frames contain polyp annotations and in total 200 labelled polyps [56]. Polyp annotation is improved by using an object-tracing- based intelligent annotation tool. Additionally, it offers 103 videos, including 861,400 frames without annotations. Therefore, unsupervised and semi-supervised methods will be able to benefit from these videos as they enrich the diversity of the data. Authors also evaluated the data using you look only once (YOLO), retina network and centre network for polyp detection.

The BKAI-IGH NeoPolyp database is an open-source colonoscopy polyp database developed and released by BK.AI, Hanoi University of Science and Technology incorporation with Institute of Gastroenterology and Hepatology (IGH), Vietnam [57]. The database has two versions. BKAI-IGH NeoPolyp-Small contains 1200 (1000 train, 200 test) images including 1000 white light imaging and 200 flexible spectral imaging colour enhancement images [57]. It consists of 2 classes of polyps namely neoplastic (red) and non-neoplastic (green). Segmentation and classification annotations are available as ground truth which are verified by experienced endoscopists at IGH, Vietnam. The NeoPolyp database is bigger and contains about 7500 polyp images in four different colour modes namely linked colour imaging, narrow band imaging, white light imaging, flexible spectral imaging colour enhancement. An additional class, ‘undefined’ polyp is also included (yellow colour) in the NeoPolyp database. The larger dataset is not publicly available yet. The application of the BKAI-IGH NeoPolyp database is polyp segmentation with more focus on incisive classification for neoplasm polyp identification.

Endotest dataset is an annotated polyp colonoscopy database, constructed with the aim to compare various polyp detection systems [58]. Endotest along with containing sequences of polyp and non-polyp frames, also includes frame- wise full length polyp colonoscopies [58]. The data was provided by 2 centres (University Hospital Ulm and Würzburg) using Olympus CV-190 endoscopy processor. It consists of 2 sets i.e., a validation dataset which contains 48 videos with 12159 polyp images and 10697 non-polyp an the second, performance dataset which has 10 full-length colonoscopies with 36482 polyp and 194416 non-polyp ones [58]. They are manually annotated.

Colonoscopic dataset is a collection of various polyp classification video datasets namely MICCAI 2017 [46], CVC colon DB [41], GLRC [44], and data collected from University of Kansas Medical Centre (KUMC). The KUMC dataset includes 80 colonoscopy videos which were then manually annotated. The motivation behind developing this dataset was to open source a colonoscopy-based poly detection database with a large number of samples so as to compare various deep learning models. Each frame is manually labelled with various polyp classes as well as locations. A total of 155 video sequences or 37,899 frames are available with bounding boxes and labelled polyp classes acting as ground truth. Moreover, the data is split into 116 training, 17 validation, and 22 test sets.

Table 2.2 Comparative analysis of the available colonoscopy datasets in this field.

Ref.	Releasing year	Size	Acquisition	Image/video format	Image resolution(pixels)	No. of patients	Ground truths	Medical validation
[53]	2021	2.393GB	Olympus Exera 195,	.jpg	384 × 288	600	Bounding	No
			160AL Olympus		to 1920 × 1080			
[49]	2017	1.2GB	endoscope H190	.jpeg	720 × 576, 1920	-	boxes	Yes
		(v1),	ScopeGuide, Olympus		× 1072			
[51]	2020	2.3GB (v2)		.jpeg,		-	Annotations	Yes
		44.1MB	ScopeGuide TM	.json(masks)	332 × 487-1920			
					× 1072			
[50]	2020			.jpeg,		-	Masks, class	Yes
		58.6GB	Olympus Europe,	.json(masks)	720 × 576, 1920			
[45]	2015		Pentax Medical	.png	× 1072	31	labels Masks,	Yes
		46.8MB	Europe					
[41]	2012			.png	384 × 288	13	class labels	Yes
[47]	2017	-		.png		44		Yes
		-	-		500 × 574		binary masks	
			-		500 × 574, 384 ×		, annotations	
[47]	2017			.png	288	-	binary masks	Yes
[42]	2014	-		.png		34	Masks, lumen	Yes
[43]	2015	-	-	.png		286	& specular	Yes
[41]	2021	-	-	.mp4	1920 × 1080	31	highlight	Yes
[52]	2020	-	Olympus, Fuji	.jpeg	1225 × 966	40	annotations	Yes
		3.2GB	OLYMPUS		1225 × 966		binary masks	
			endoscopes		768 × 576		binary masks	
[54]	2020		Olympus	.jpg	854 × 480, 1920	1600	binary masks	Yes
					× 1080			
		-	endoscopes				binary masks	
[59]	2022			.jpg		-	binary masks,	Yes
		17.1GB	(CF-H190L & CF-		256 × 256	13	class labels	
[48]	2021		HQ190L)	.png		1731		Yes
[55]	2021	1.1GB	Olympus PCF-	.jpg	768 × 576		binary masks	Yes
		-	H290DI, FUJIFLIM			-		
[56]	2021		EC-530wm	.jpg	500 × 574		class labels,	No
[60]	2022	24.28GB	-	.jpeg	192 × 192	-	annotations	Yes
[58]	2022	375.42 MiB		.mp4, .jpg		-	-	No
[44]	2016	513.6MB	-	.mp4	-		Bounding boxes	Yes
		-	CF-HQ290ZI and		1280 × 995			
			CF-H290ECI;		-	-	Annotations	
[61]	2021		Olympus	.mp4	768 × 576		Class labels,	No
		2.3GB	-				masks	
			-				Bounding	
			Olympus CV-		various resolutions		boxes	
			170,190 Olympus				Histopathology	
			ExeraCV180				and camera	
			and Olympus Exera-				calibration	
			CV190				masks, class	
							labels	
			-					

Table 2.2 presents existing available datasets of colonoscopy polyps for AI enabled techniques. They are compared on the basis of their releasing year, size of the data,

mode of acquisition, type of image or video format, presence of polyps and non-polyps, medical validation and presence of ground truths in the datasets. The first colonoscopy dataset released in the year 2012, namely CVC ColonDB and CVC-PolypHD have paved a way towards scientific research in early diagnosis of GI tract related diseases and effective management through robust AI techniques. SUN colonoscopy has recorded from maximum no. of patients i.e., 1731 out of which 1405 were eligible for the study. Only CP-CHILD dataset has focused on collecting colonoscopy data of children suffering from colorectal disease up-to 18 years of age. SUN and Endotest data contain maximum polyp and non-polyp frames amongst the twenty-one discussed datasets. Most datasets have been released with binary masks which play an important role in segmentation tasks. Fewer, recently released datasets like Polypgen have provided sequential frames and most datasets contain de-identified videos or its frames for analysis.

Available datasets still lack consideration of real-time medical settings. Real- time scenarios contain glare and artifacts in the colonoscopic video frames due to the coaxial arrangement of the lens and light source leads. The brightness and contrast of recorded video frames may not be of high-resolution due to varied reasons and depends on the geometry of the tissue area and liquid cleansing, and medical resource, cost, camera settings. Blurred frames are also recorded due to the movement of the sensor inside the organ cavity. The captured tissue images vary significantly during the contraction of the muscle fibres of the organ due to the peristalsis movement in the colon and rectum region of the body. Mode of acquisition also brings variations in the recording. It is clear from Table 2.2 that the image resolution varies significantly in all the available datasets. Several datasets like Polypgen, Kvasir SEG, etc, contain medical information on the left side of the video frames and contain a black box in the extreme left corner. Such frames require data cleaning methods before acting as input to AI pipelines. Cleansing score analysis of the endoscopic video is another important area of research which is still emerging.

Now we discuss the dataset available in VCE. High-quality, open-access and free VCE data act as a catalyst for on-going state-of-the-art AI research works in management of various GI tract related diseases such as Crohn's disease, colorectal cancer, GI bleeding, motility disorders, celiac disease, inflammation, polyps, and hookworms etc [24], [62], [63], [64]. Researchers have focused on different problems related to VCE technology like anomaly detection, video summarization, noise and artefact removal, and processing etc wherein each problem requires a unique, and high-quality VCE data for analysis [20], [24], [65], [66], [67]. Table 2.3 shows a comparative analysis between the existing, available VCE datasets for research use.

KID project VCE data was originally launched as an internet-based digital video atlas for VCE in 2017 [68]. It is the first dataset which is actively utilized by researchers and is available for medical image analysis in VCE research and remains the benchmark dataset since 2017 [68]. It contains more than 2500 annotated image and 47 videos and was collected from six centres. All the videos were acquired using MiroCam (IntroMedic Co, Seoul, Korea) capsule endoscope. KID data is bifurcated into two datasets, three videos and its parts. Dataset 1 consists of 77 VCE images which has various anomalies like angioectasia, apthae, chylous cysts, polypoid lesions, villous oedema, bleeding, lymphangiectasias, ulcers and stenoses. Dataset 2 consists of 2371 VCE images. Several polyploids, vascular and, inflammatory lesions observed in small bowel region are included in this dataset along with normal images from the GI tract. It was developed with a high standard image quality, protocols and standard annotations for state-of-the-art research purposes.

Kvasir-Capsule is currently the largest and most recent, diverse, publicly available VCE dataset which contains 4,741,504 normal and abnormal image frames extracted from 117 anonymous videos [65]. It was collected from Norwegian Hospital using Olympus EC-S10 endocapsule. It was released in 2020 on OSF home, a free website to share research works. The findings are divided into two categories i.e., anatomy and luminal findings. 47,238 video frames are labelled and contains a bounding box for each frame. Rest of the frames are unlabelled. Both video and images which are labelled and unlabelled are available for download. All the annotations have been medically verified by four specialized hospital doctors. However, no cleansing grades during bowel preparation, class labels and related medical information to assess the grade of cleanliness has been done.

Gastrolab is an image gallery-based website which contains raw images, and video segments of different type of endoscopies of varied locations, patients and, GI tract diseases and their types [69]. It is managed by science photo library. Red Lesion Endo-scopy Dataset is the first publicly available VCE dataset which contains red lesions like angioectasias, angiodysplasias, and bleeding only. It is freely available on

Table 2.3 Comparative analysis between the existing, available VCE datasets for research use.

Name of the CE dataset	**Data collection place and year of release**	**No. of classes and VCE images**	**Limitations/ Remarks**
KID	Royal Infirmary of Edinburgh, United Kingdom and 2017	8 and 683 images	Lack cleansing score, sequential frames and multi-label medical information. Benchmark dataset in AI WCE research.
Kvasir-Capsule	Department of Medicine, Bærum Hospital, Vestre Viken Hospital Trust, Norway and 2020	14 and 47,41,504 images	Lack cleansing score, and multi-label medical information. Presently largest WCE data
Gastrolab	Multiple labs across globe and -	4 and -	Raw VCE video frames and video segments of crohn ulcerations, normal GI tract, hyperplasia, and whipples.
Red Lesion Endoscopy (MICCAI 2017)	Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Portugal and 2018	2 and 3895 images	Contains only red lesions. Medical annotation is done for Set A only.
Annotated bleeding (Farah deeba)	- and 2016	2 and 50 images	Contains only bleeding and non-bleeding WCE images which may not be medically verified.
Crohn IPI	Nantes University Hospital and 2020	7 and 3498 images	Focused on one disease i.e., Crohn’s disease
EndoSLAM	Multiple labs in UK, USA And 2020	- and 3294 images	Developed for depth estimation and Monocular Visual Odometry, contains Tumor frames.
VRCaps	Multiple labs in UK, USA and 2021	-	- (Artificial data synthesis of CE technology)
CE Cleanliness	Hospital Universitari i Politècnic La Fe,Valencia and 2020	2 and 1417 images	Lacks anomaly information, class labels of anomalies, medically validated cleansing scores

INESCTEC data repository since 2018 and contains two sets. Set 1 and 2 consists ofabout 3295 non-sequential and 600 sequential frames respectively with manually annotated masks. The videos were recorded using different cameras like MiroCam, Pill-Cam SB1, SB2 and SB3. Annotated bleeding dataset was released in 2016 for automatic segmentation of 50 bleeding and non-bleeding VCE images [70]. The ground truth images are also available. No other information is available for this dataset. Crohn IPI dataset contains 3498 WCE image frames from different pathological findings and normal images collected using PillCam3 at Nantes University Hospital [71]. All the abnormal videos were taken from patients suffering from Crohn's disease. The annotations were medically verified by three independent medical experts. Various anomalies like ulcers, lesions, erythema, edema, and stenosis etc. The data is available on request from the authors.

EndoSLAM dataset contains different types of endoscopic recordings of porcine GI tract organs, synthetic data generated through VR caps and phantom recordings of the colon and other GI tract organs with computed tomography scan ground truth [72]. The recordings have been acquired using conventional endoscopes like Olympus colonoscope and newer WCE cameras like MicroCam, Pillcam colon, high- and low-resolution cameras. 42,700 image frames distributed in 35 sub-datasets have been developed which aims to provide 6D depth pose, 3D ground truths, image frames from different cameras views, angles and light conditions. A virtual environment has also been released on the similar area of interest known as VR caps [73]. Both of them are available freely on Github.

CE cleanliness is the only dataset which is available for assessing the grade of cleanlinesss [74]. It was released in 2020 and acquired at Hospital Universitari i Politècnic La Fe from Valencia, using Pillcam SB 3 system. 563 individual frames of 576×576 pixels were extracted from 35 different CE videos during patient procedures and considered as training set. 854 additional frames of 576×576 pixels were extracted from 30 additional CE videos of different patients for development of validation set. The dataset has been built to locate and quantify the intestinal content in a CE procedure wherein the extracted frames have been cut into patches of 64×64 pixels, with a step size of 32 pixels which are class labelled as dirty or clean. No other medical information related to the findings of the CE and their class labels has been mentioned in the dataset.

2.3 Abnormality Detection Methodologies in Endoscopy

Abnormality detection in endoscopy involves identifying deviations from normal tissue appearance during internal examinations of organs like the GI tract [18], [24], [26], [75]. Utilizing specialized cameras, this diagnostic technique helps detect anomalies such as polyps, lesions, or inflammation. Computer-aided systems, employing image analysis and machine learning, assist in recognizing subtle abnormalities, enhancing diagnostic accuracy. Early detection through endoscopic abnormality recognition is crucial for timely intervention, enabling effective treatment of conditions like colorectal cancer [11], [12], [76], [77], [78]. This approach combines medical expertise with technology, contributing to improved patient outcomes by ensuring the early identification and management of abnormalities within the visual field of endoscopic procedures [79], [80].

In this thesis, we will focus on a specific abnormality known as polyp. A polyp is an abnormal growth of tissue that protrudes from the mucous membrane (inner lining) of an organ [81], [82], [83]. Polyps can occur in various parts of the body, including the colon, stomach, nasal passages, uterus, and other organs. In the context of gastrointestinal health, such as colonoscopy or endoscopy procedures, the term ‘polyp’ commonly refers to growths in the lining of the colon or rectum.

Polyps vary in size, shape, and appearance, and they can be classified into different types based on their characteristics [83]. Some common types of colorectal polyps include adenomatous polyps, hyperplastic polyps, and serrated polyps [83]. Adenomatous polyps, in particular, are of concern due to their association with an increased risk of colorectal cancer [82], [83]. It's important for individuals, especially those at higher risk, to undergo regular screenings and consult with healthcare professionals to monitor and manage the presence of polyps for preventive health measures.

One of the early studies on automatic detection of polyps using colonoscopy frames focused on identifying polyp region through texture-based feature extraction followed by classification of support vector machine (SVM) to identify the position of the polyp [84]. The authors developed sub-frames of the main frame and labelled them normal region and polyp region to develop a training dataset. The frame resolution was of 378×254. Texture features were calculated for each of the sub-image and then classification task was performed to identify the position of the polyp. The authors achieved a sensitivity up-to 86.2% on their test dataset of seventy-four colonoscopy frames. The simulation codes were run on a 1.83GHz Intel Centrino Duo CPU and 2GB RAM computer. A similar study achieved a region over curve (ROC) curve up-to 0.96 using texture-based features [85]. The authors pointed out that the major drawback of automating the process is the lack of the high-resolution data in this field. So, the authors collected a high-resolution colonoscopy video with resolution of 1920×1080 and processed it for automatic detection of polyps. This way the sub-images were of higher resolution and more precise features were extracted from it. In continuation to the study, an automatic polyp region segmentation using watershed algorithm and ellipse segmentation was proposed [86]. The authors showed that texture-based features had a limitation over frames which contain multiple polyps and demanded a fixed-sized analytical window to be cut out as a sub-image. The elliptical shape-based algorithm covered small and large polyps and achieved a sensitivity and specificity of 93% and 98% respectively.

Similar studies surfaced from 2010-2015 wherein the researchers utilized hand-crafted features which were colour-based, statistical-based, texture-based, special-information based to identify polyp and normal tissue using different machine learning classifiers like SVM, random forest (RF), k-Nearest Neighbours (KNN), linear regression, logistic regression (LR), etc. Most studies consisted of splitting the frames into a 50-50 manner to train and test their AI models over colonoscopy frames ranging from 50-500 [41], [45], [84], [85], [86], [87], [88]. MICCAI is focused on advancing research in MIA and AI, launched its first biomedical challenge on polyp recognition. State-of-the-art datasets were released wherein CVC Colon DB was also released. Interested researchers and industry professionals participated in these challenges from 2015-2017 to work on the problem of automatic detection of polyps in colonoscopy frames. However, the need of robust datasets which contain diverse polyp frames with varied size, type, and texture remained a research gap. The dataset released in the challenges were not accessible for public research use at the time. In the subsequent years, partial datasets were released for public research use. The existing selection bias problem in these datasets remained an un-looked area. It was only realised in a study published in a medical journal wherein the doctors showed that the AI model interpreted a frame as polyp which was completely resected by the doctors during polypectomy [89]. The study showed the biasness of the AI model and showed the need to include non-polyp frames, incomplete polyp resection-based frames, and unwanted frames which were discarded in the existing datasets.

With the advent of ensembles and deep learning models, this field too experienced the shift. Different deep learning models YOLO, single shot detector (SSD), custom convolutional neural networks (CNN), faster CNN, u network etc were implemented in this field [26], [75], [90], [91]. Some of the studies have been discussed here.

Wang *et al.* [92] conducted the first prospective randomized control study to automatically detect polyps using segment network, a deep learning architecture. The architecture was trained and tested on private data collected from the Endoscopy Center of the Sichuan Provincial People's Hospital, China. The deep learning architecture was able to detect about 81.16% polyps present in 292 colonoscopy videos in comparison to the human operator who detected about 50.1% polyps present in 293 colonoscopy videos. The false positives were reported for images with bubbles, faeces, undigested debris, wrinkled mucosa, rounded drug capsules, and local inflammation in the area. Zhang *et al.* [93] proposed an SSD Global Pooling Network (GPNet) for the automatic detection of gastric polyps, in real-time, using private data from Sir Run Run Shaw Hospital, China. The architecture consisted of feature pyramids along with a backbone of visual geometric group (VGG) 16 to automatically extract unique features of polyps in colonoscopy frames.

Hsu *et al.* [94] proposed a CNN network consisting of three CNN layers followed by a max pooling layer for polyp detection and classification with the input of grey-scaled polyp frames. CVC-Clinic and privately collected polyp video frames were used to train the network. Soons *et al.* [95] proposed a real-time polyp detection using the

DISCOVERY computer-aided detection (CAD) system on privately collected data from Europe. The authors have reported about twenty percent false positives rate in eighty-one colonoscopy videos.

Jha *et al.* [96] proposed a deep learning architecture named colon segment network to automatically detect, localize, and segment polyp frames using the Kvasir-SEG dataset. To tackle the high number of false positives found in neighbouring frames, Qadir *et al.* [97] proposed a CNN-based object detector network with a False Positive (FP) reduction unit. The authors trained and tested their network on CVC-Clinic, ASU-Mayo Clinic, and CVC-Clinic-Video DB datasets. Rodrıguez *et al.* [98] used a fine-tuned YOLO version three architecture to detect polyps in real-time using privately collected data from Chuo University, Japan. The pipeline was developed and evaluated using a Compi-based framework.

Nehme *et al.* [99] evaluated the effectiveness of the first Food and Drug Administration (FDA) approved CAD device to automatically detect adenomas in daily clinical practice. The authors concluded that the CAD device did not improve adenoma detection in the clinical setup. It was found prone to a high number of false positives, and a high level of distraction. It was also susceptible to prolonged procedure time. Krenzer *et al.* [100] proposed a deep learning architecture based on YOLO called the ENDOMIND-Advanced to detect polyps in real time. The authors merged seven open-source datasets (out of twenty-three) and one privately collected data from Germany to train, test, and validate their deep learning architecture.

Sanchez *et al.* [75] pointed out that the existing state-of-the-art experimental studies have focused on polyp detection and classification with deep learning architectures using private datasets. Most experimental studies included in their review study had not tested their architectures for human-unaltered video datasets due to a lack of open-source datasets. The same may be observed through the studies discussed above. Additionally, the above discussion reveals that there is still scope for improvement in the detection of abnormality specifically for polyps, in routine clinical and real-time use.

2.4 Cleanliness Assessment Methodologies in Endoscopy

Cleanliness assessment in endoscopy involves two major aspects namely the quantification of chemicals which are given to a patient to clean the tract of the patient before endoscopy, and scoring methods which may analyse the endoscopy video for their cleanliness post the endoscopy. The first aspect is decided through the state of the patient, their allergies, resistance to any drug, experience of the endoscopist, and standard rules followed at the hospital. The second aspect is still state-of-the-art as most scoring methods are not standardized for video analysis and assessed through frame-by-frame analysis, and have not been tried and evaluated for large, and multi-centre clinical trials. In VCE, purgatives like polyethylene glycol, sodium phosphate and simethicone etc are given to the patient to prepare their bowel. Several studies have been on-going to check which purgative is best in nature and whether bowel preparation before VCE enhances the diagnostic yield of the VCE or not. Contradicting results have been reported over the years. Nevertheless, the manufacturers of the VCE suggest fasting of at least 6-12 hours before the procedure, and in-take of liquid diet before and after the procedures for 2 days. In this thesis, we have hypothesized that the assessment of cleanliness in VCE is crucial for ensuring optimal visualization and, consequently, accurate diagnosis or the so-called diagnostic yield of the VCE. Cleanliness metrics, the scoring system must play a pivotal role in maintaining the quality of VCE examinations, allowing gastroenterologists to make informed decisions based on clear and unobstructed images.

Now we will discuss the state-of-the-art methodologies surfacing for automatic cleanliness assessment in VCE. One of the initial studies appeared in the year 2015 wherein Klein *et al.* [101] designed and validated a computed small bowel preparation score based on the pixels in the colour bar of red-green-blue (RGB) images. The authors categorized the images as ‘adequate’ and ‘in-adequate’ to perform CE image classification and correctly classified 71 CE videos out of total 85 CE videos. Pietri *et al.* [102] focused on developing a computer-aided system based on four different statistical features namely grey level correlation matrix (GLCM), fractal dimension features and Hough transform to evaluate the abundance of bubbles in CE. 400 still frame was categorized as ‘scarce in bubbles’ or ‘abundant in bubbles’ based on the percentage presence of bubbles observed by the physician in the CE frame. A sensitivity up-to 95.79% and 94.74% was achieved for specified features on the validation set.

Abou Ali *et al.* [103] developed a computer-aided assessment of cleansing (CAC) score to inspect the effectiveness of evaluation of quality for CE still frames. The authors used channels of colour intensities of the RGB model and extracted it for each frame. A SB-CAC score cut-off of 1.6 validated a sensitivity of 91.3 % and a specificity of 94.7 % for 228 still frames categorized as 'adequate' and 'inadequate'. Oumrani *et al.* [104] proposed an automatic rapid tool for assessing mucosal visualization quality of still CE images using colour intensity ratio, brightness index and GLCM features and random forest classifier. 600 normal still CE frames were extracted and evaluated through 10-point assessment grid. The combination of the mentioned features produced a sensitivity up-to 90%.

Noorda *et al.* [74] developed a CNN architecture with light weight and reduced trainable parameters to automatically evaluate the degree of cleanliness in CE on an intuitive scale such as 'clean' or 'dirty'. In order to locate and quantify the intestinal content, the authors developed patches from the extracted VCE frames. Based on the 5-fold cross validation performed on 35 video patch frames for training set and 30 different videos for validation set, an average classification accuracy up-to 95.23% has been reported.

Nam *et al.* [105] developed a deep learning-based software to calculate cleansing score in CE. They used 700 images per each cleansing score and implemented an existing deep learning model named Inception residual network version 2 (Inception ResNet V2). Abnormality such as polyp, ulcer, bleeding etc., found in the frame were not considered in the training and test set. The top-1 and top-2 accuracies achieved by the deep learning network were 69.4% and 91.2%, respectively. Based on the classification, scores were manually assigned and compared with physician's decision to check the efficacy of the network. Similar work by the authors used Generic CNN consisting of five convolutional and max pooling layers with one full connected layer for 4,00,000 still frames categorized into score 1-5 depending on mucosal visibility.

The proposed network achieved an accuracy up-to 93% on 120 test set frames and misclassification rate up-to 24.7% on 51,380 separate set of CE frames. A VGG16 neural network based automatic SB cleanliness scoring was proposed using 600 normal still CE frames [106]. They were categorized as ‘adequate’ and ‘inadequate’ based on 10-point scale. The authors reported an accuracy up-to 89.7%. It may be noticed that only one dataset is presently available which doesn’t contain medical-based scoring information and contains binary labels. Most studies have been done using private datasets with two labels. To the best of our knowledge, no study has been to automate an existing medical score for VCE cleanliness assessment. Additionally, real-time studies have not been tried further suggesting a scope of improvement in this field.

2.5 Multi-label Classification in Endoscopy

Multi-label classification in endoscopy involves categorizing endoscopy videos or frames into multiple classes or labels, indicating the presence of various abnormalities or conditions or scores simultaneously. This approach is essential in gastroenterology, where endoscopy procedures may reveal multiple pathologies in a single examination, such as identifying polyps, inflammation, or ulcers. Advanced machine learning algorithms may analyze endoscopic imagery to classify and label diverse abnormalities, enabling comprehensive diagnosis and treatment planning. By addressing the complexity of detecting multiple conditions in a single endoscopic session, multi-label classification may contribute to be more accurate and provide holistic assessments of patients' GI health during endoscopy procedures. Upon conducting a thorough search on engineering village, a famous literature search website by Elsevier, google scholar, PubMed, IEEE explorer, and science direct for years 1950-2024, we found that multi-label classification in endoscopy is still an emerging area. Primary reasons behind this are the already existing lack of data in endoscopy for AI-enabled techniques. Scarce data is available for binary labels. Multi-label data collection demands collaboration with experienced gastroenterologists and understanding of the medical problem. Herein, we discuss the existing studies in multi-label classification for endoscopy.

Vasilakakis *et al.* [107] investigated the semantics of a CE video content such as mucosal tissues, hole of the lumen, bubbles and debris through a weakly supervised framework. This framework consisted of salient point detection algorithm, bag-of-word representation and multi-label classification using support vector machine classifier. The authors extended their work and developed a multi-label classification method for five categories namely abnormal, debris, bubble and lumen hole using bag-of-words approach on CIE-Lab converted CE RGB frames and a convolutional neural network architecture enabling multi-scale feature extraction (MM-CNN). The experimental work used KID data wherein multi-labels were developed using Ratsnake software. Keywords like ‘abnormal’, ‘debris’, ‘bubble’ and ‘lumen hole’ were added to a particular CE frame containing a lesion abnormality like ‘polyp’, ‘ulcer’, ‘bleeding’ etc. It was done to finally check whether the CE frame has existence of abnormalities or not, the existence of debris or not, etc. They did not consider dependencies between labels and associated cleansing score. The developed method achieved an area under curve (AUC) score up-to 0.94, 0.91 and 0.85 for debris, bubble and lumen hole classes.

Park and Lee [108] proposed a class-labelling method that can be used to design a learning model by constructing a knowledge model focused on main lesions defined in standard terminologies for CE such as minimal standard terminology and CE structured terminology The knowledge model considers the anatomy of the GI tract and findings in CE. Three major class labels namely normal, abnormal and discriminative class have been given wherein the normal class labels are further distinguished based on the bubbles, wrinkles, location of the capsule and discriminative classes contains frames due to low power, transmission or reception problems, and a large amount of foam. The special cases have been analyzed by developing clusters of colour similarity through k-means algorithm. The supra- and sub-classes have been made using the concept of ontology. The authors conducted a classification task to distinguish the different organs using a generic CNN and achieved an accuracy up-to 33.5%.

Mohammed *et al.* [109] developed a pathology-sensitive abnormality detection through CNN pipelines, attention, residual long short-term memory (LSTM), and selfsupervision sub module for colon diseases in VCE data. They developed VCE dataset (presently private) containing 455 short video segments with 28,304 frames and 14 classes of colorectal diseases. The classes consisted of abnormalities such as erosions, debris, diverticulosis, erythema, granularity, haemorrhage, inflammation, edema, angioectasia, polyp, pseudo polyp, tumour and ulceration. A total of 227 and video were considered for training and testing the proposed AI method. The authors performed video and frame-level prediction and achieved a precision, recall, F1-score and specificity up-to 61.6%, 54.6%, 55.1%, and 95.1% respectively.

2.6 Chapter Summery and Gaps in the Study

This chapter delves into the critical advancements and challenges in colonoscopy and VCE, specifically focusing on two pivotal aspects: automatic abnormality detection, with a primary emphasis on polyps, and cleanliness assessment for maintaining the diagnostic yield and video quality of VCE. The key points of the chapter include:

- Introduction to the evolution of computer-aided systems in colonoscopy and VCE for a reliable and fast automated analysis which saves the time and effort of gastroenterologists.
- Exploration of colonoscopy and VCE datasets for Ai-enabled techniques.
- Emphasis on existing methodologies for automatic polyp detection and cleanliness assessment in VCE.
- Advancements in multi-label classification of endoscopy data.

Pertaining to the literature discussed in the above sections, following are the research gaps:

- Lack of high-quality, large endoscopy dataset for abnormality detection and cleanliness assessment.
- Need for robust and effective methodologies to detect abnormalities in endoscopy.
- Need for precise and fast methodologies to assess the cleanliness in endoscopy.
- Need of multi-label classification in endoscopy.

2.7 Conclusion and Future Scope

In this chapter, we discussed the existing colonoscopy and VCE datasets and methodologies which may be utilized to tackle the problem in automation of endoscopy analysis, specifically for automatic polyp detection and cleanliness assessment in VCE. In both the problems, the traditional approaches initiated with handcrafted features such as the texture, colour, and statistical abstractions and application of machine learning algorithms. The approaches have slowly evolved to automated feature extraction through deep learning algorithms. There is a need of development of large dataset and methodologies in both the field for advancement in this field.

The above discussed research gaps have acted as a source of motivation for the development of new methodologies in this field. This thesis will focus on developing and releasing two new datasets namely gastrointestinal polyp dataset and KODA dataset to advance research and simulations for automatic polyp detection in colonoscopy and cleanliness assessment in VCE. Two new methodologies will be discussed in detail to combat the research gaps of need of robust and effective methodologies to detect abnormalities, assessment of cleanliness and multi-label classification in endoscopy primarily VCE in the Chapter 3, 4, and 5.

CHAPTER 3

DEVELOPMENT OF DEEP LEARNING ARCHITECTURE FOR ABNORMALITY DETECTION IN ENDOSCOPY

3.1 Introduction

The incidence rate of GI and liver diseases has significantly increased, especially in developing countries like India, China, and Japan, due to the increased use of antibiotics, and changes in environmental conditions, health, and diet [3]. Colorectal cancer is one such common GI disease that is characterized by persistent changes in the bowel, bleeding in the rectal region, colorectal polyps (CP), and discomfort in the abdomen [76]. The cumulative risk of colon cancer developing for an un-removed and un-treated polyp is about 24% at 20 years after its diagnosis [83]. Colonoscopy, an endoscopic method has been suggested as an early diagnostic method to remove and treat these colon and rectum polyps [76]. The procedure takes about 30-60 minutes which records for an adjustable frame rate of up to 4-30 frames per second [17]. Manual analysis of large colonoscopy frames for a single patient is not only time-consuming but also repetitive for an experienced gastroenterologist [17], [110]. Considering the low doctor-to-patient ratio, and lack of resources, various machine learning, deep learning, and transfer learning pipelines have been proposed to introduce automation in the detection of CP while reducing the false positive rate and eventually, aid in the burden on skilled gastroenterologists [111], [112]. Several handcrafted features such as scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), colour space statistical features, etc., have been proposed in the past [113].

3.2 Related Works

Deep learning architectures such as faster-residual CNN, single shot multi-box detector, YOLO, Inception version 3 (V3), VGG 16, Resnet V2, and Resnet-150, etc., have been frequently used for CP detection, segmentation, and classification [26], [29], [75], [90], [91]. Hsu *et al.* [94] proposed a CNN consisting of three CNN layers followed by a max pooling layer for CP detection and classification with the input of gray-scaled CP frames. CVC-Clinic and privately collected CP video frames were used to train the network. The study also inferred through hyper-parameter tuning that the accuracy of the proposed network decreases dramatically when images of polyps are smaller than 1600 pixels. Another study proposed an architecture with four convolutional layers and max pooling layers followed by two fully connected layers to detect CPs. Rahim *et al.* [114] proposed a sixteen-layered CNN with MISH activation function instead of rectified linear unit activation (ReLU) for automatic polyp detection. Two research works have been carried out in interpreting the deep learning architectures. The first study, by Wang *et al.* [92] focused on CP classification, segmentation, and detection using Shapley additive explanations and performance predictors. The second study checked the uncertainty in CNN while performing colorectal polyp semantic segmentation using the uncertainty method called the Monte Carlo Guided Backpropagation [115]. Kliegis *et al.* [89] investigated the famous, pharmaceutically approved, CAD-Eye AI system (Fujifilm Europe) for polyp detection and characterization after completing endoscopic resection of colon adenomas. The AI system poorly misidentified all the seventeen resection sites and its exposed normal sub-mucosal tissue as polyp regions showing that the system should be trained and tested for occluded frames to avoid such misidentifications. Excellent review developed by Rodríguez *et al.* [26], [98] pointed out that the current research works in this field have focused on polyp detection and classification using deep neural networks (DNN) but still lacks the testing on human-unaltered video datasets, larger, and freely, available datasets like ImageNet for computer vision related tasks, with variable CP and non-CP frames. Hence, there is still room for improvement in terms of extensive test set analysis, occlusion testing, and hyper-parameter tuning leading to the development of trustable and explainable AI pipelines for automatic CP diagnosis.

3.3 Problem Statement

With the rise in the use of deep learning architectures, the concept of `end-to-end' learning has become popular wherein the feature extracting and pre-processing steps have merged in the deep learning pipeline and are often performed `on the fly' to save computational time and memory [39]. Pharmaceutically approved deep learning systems such as GI genius (Medtronic Plc, Dublin, Ireland), DISCOVERY (PENTAX Medical, Tokyo, Japan), Endo-BRAIN-EYE (Olympus Corporation), and CAD-EYE (Fujifilm, Tokyo, Japan) have motivated researchers to develop robust architectures which are able to detect even the minutest CPs of dimension less than 2 mm [26], [29], [75], [90], [91]. However, due to the complexity and black-box nature of these architectures, developing trust among the user (clinical staff and doctors) is presently difficult. Recent reviews done in this field have shown the lack of separate validation and test set analysis due to the unavailability of open-source datasets in video or still, image form [26], [29], [75], [90], [91]. It further has led to poor generalizability and over fitting of existing state-of-the-art deep learning architectures [26], [29], [75]. In continuation, it has also hindered the analysis of sequential and non-sequential frames in colonoscopy videos.

These architectures consist of various hyper-parameters like kernel size, weights, kernel initializer, learning rate, optimizer, batch size, scaling, data augmentation methods, etc. Fine-tuning of these hyper-parameters and their block-by-block removal through ablation experiments are vital and can be further explained and tuned as per the requirements of health officials. Architecture explainability and transparency are important, that focus on how deep learning architectures associate a certain CP video frame with its class label and further depict factors that influence this prediction. For instance, a recent study by Wickstrøm *et al.* [115] investigated the uncertainty in the prediction of U-Net and SegNet, CNN based segmentation architectures for CP diagnosis. Another study conducted a survey of intrinsic and extrinsic explainable methods for health officials in the gastroenterology department and reported the explainability of Gradient-weighted Class Activation Mapping (Grad-CAM) to be the most efficient and preferred amongst doctors in comparison to Shapley Additive exPlanations (SHAP) values [92], [116].

Owing to the above-discussed research gaps in this field, the present work proposes an explainable, end-to-end, automatic colorectal polyp diagnosis architecture called `Window bAsed Detection afTer Mixed Convolutions Polyp Identification (WADT-MCPI)'. The main contributions of the work are concluded as follows:

- The work proposes an end-to-end deep learning architecture called `Window bAsed Detection afTer Mixed Convolutions Polyp Identification (WADT-MCPI)' for an automatic colorectal polyp diagnosis under colonoscopy.

- Explainability of the proposed architecture through feature mapping, CAM, and ablation studies.
- Development and release of a test set named `Gastrointestinal atlas-Colon Polyp' for a separate test or validation set analysis in this field.
- Extensive test set analysis on a variety of sequential and non-sequential colonoscopic frames while achieving an average test set accuracy up to 93%.
- The work also shows the behavior of the proposed architectures for various hyper-parameters in deep learning architectures, the developed occlusive frames and its comparison with the vanilla Inception v3 architecture.

3.4 Methodology

This section details the dataset preparation, augmentation, and experimental settings to develop the proposed architecture for automatic CP detection in colonoscopy frames. Table 3.1 shows the no. of CP and non-CP frames present in each of the datasets considered while training, validating, and testing the proposed architecture.

Table 3.1 Details of the chosen dataset.

Name of dataset	No. of polyps	No. of non-polyps	Total no. of frames
(a) Training data			
Kvasir	400	800	1200
Etis-Larib	157	-	157
CVC-Colon	299	-	299
(b) Validation data			
Kvasir	100	200	300
Etis-Larib	39	-	39
CVC-Colon	76	-	76
(c) Testing data			
Test 0	150	-	150
Test 1	256	-	256
Test 2	103	23	126
Test 3	10	13	23

The detailed specification of the chosen datasets namely Etis-Larib, CVC-Colon, and

Kvasir v1 is presented elsewhere [40]. The details of the proposed test set named ‘Gastrointestinal atlas-Colon Polyp’ are mentioned in the subsequent sections.

3.4.1 Data preparation and augmentation

Etis-Larib, CVC-Colon, and Kvasir v1 data were considered such that the training comprised 80% of the data (1656 images) and the remaining 20% formed the validation set (415 images). 1000 images from the ‘normal cecum’, ‘normal pylorous’, and ‘normal z-line’ of the Kvasir v1 data were considered to balance the non-polyp class. Both the training and validation set underwent data augmentation techniques namely shearing, zoom re-scaling, and horizontal flipping to increase the size of the dataset and introduce variability in the images since polyp size varies in a real-time setting. Other parameters like ‘channel_shift_range’ and ‘fill_mode’ were not used because they did not prove to bring many noticeable changes in the dataset. After data augmentation, a manual check was done to ensure the dis-similarity of the frames and rule out the chances of data leakage in the proposed work. Four different test sets were considered for testing the proposed architecture. Figure 3.1 depicts the variety of CP and non-CP frames considered in this work. Testing set 0 consists of the developed frames from Gastrointestinal atlas-Colon Polyp. In testing set 1, unique, random polyp frames extracted from the recently released Polyp-Gen data (Figure 3.1 Testing data (i)) were considered. Mixed, random frames collected from Polyp-Gen, the proposed test set (Figure 3.1 Testing data (g)), and the Gastro-lab website were considered in Test set 2 (Figure 3.1 Testing data (h)). Developed frames for occlusion testing were considered as Test set 3 (Figure 3.1 Testing data (j)).

The gastrointestinal atlas-Colon Polyp dataset contains seven videos. The first video is about 71-year-old man who underwent a colonoscopy for a dyspeptic syndrome containing large, ulcerated polyp with a wide pedicle and uneven surface of intra-mucosal cancer foci. Due to the patient's anticoagulation, it was advised to postpone the polypectomy till 15 days. The second video is about Recto-Sigmoid junction adenomas with chicken skin mucosa. It is associated with colonic neoplasms which are endoscopic allies and histologically aberrant. Colonic chicken skin mucosa is an endoscopic condition that develops in the lamina propria of the mucosa next to colonic tumors because of fat build up in macrophages. The pathophysiological consequences of the presence of microvilli that resembled those in the small intestine

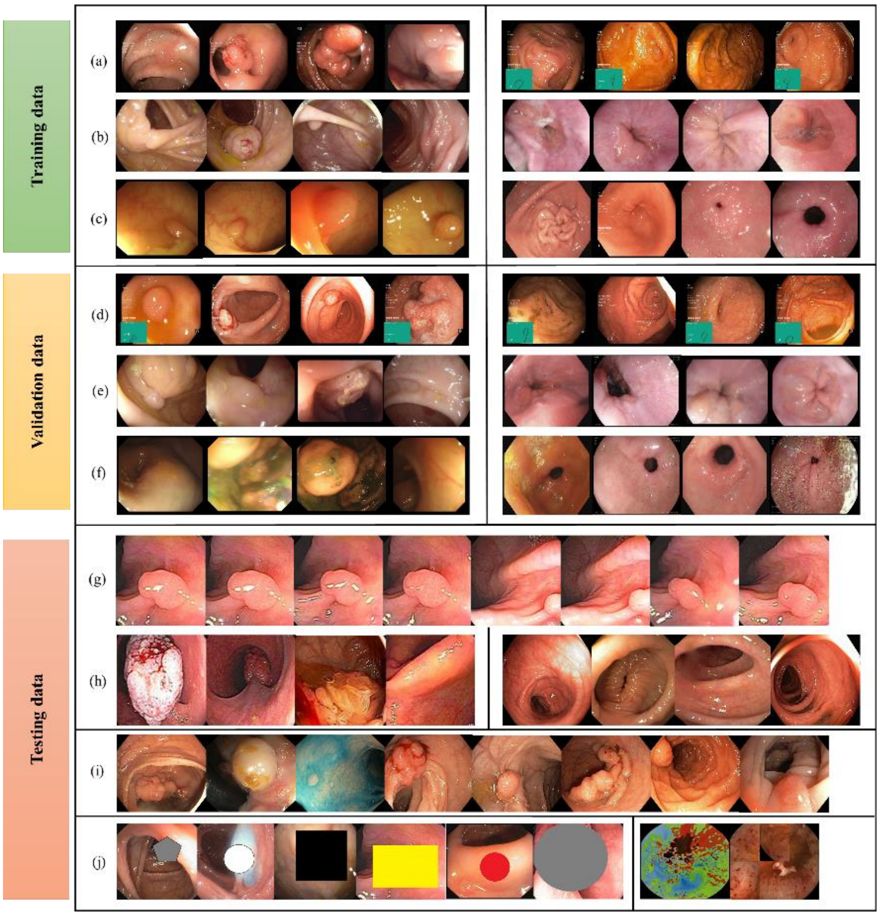


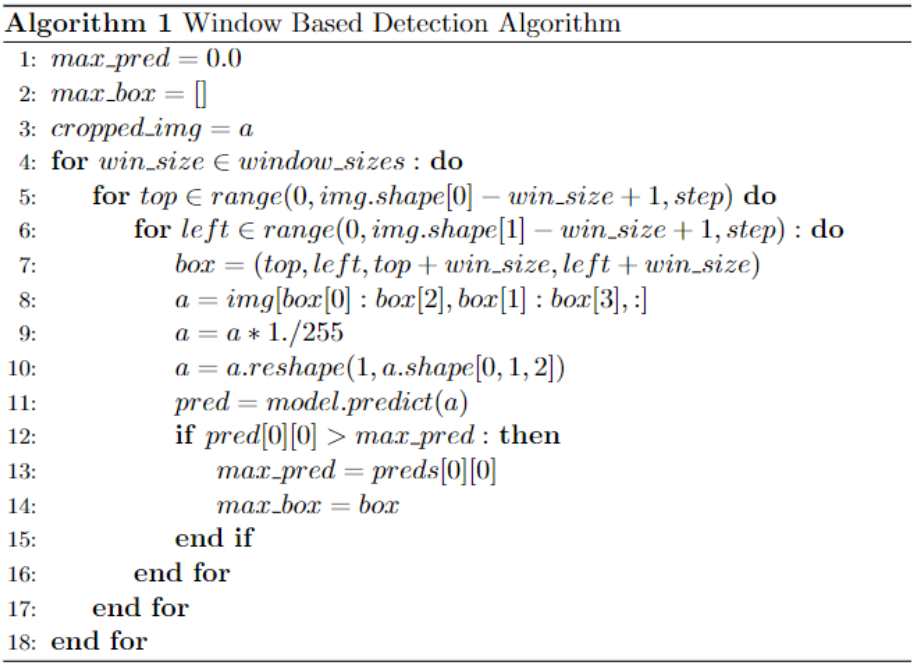
Figure 3.1 Examples of the colorectal polyp and non-polyp frames considered for training, validating, and testing the proposed architecture. Train polyp data (a) Kvasir v1, (b) Etis-Larib, (c) CVC-Colon; Validation polyp data (d) Kvasir v1, (e) Etis-Larib, (f) CVC-Colon; Testing data (g) Proposed Gastrointestinal atlas-Colon Polyp test data, (h) Mixture (i) Polyp-Gen data, (j) Developed frames for occlusion testing. The non-polyp frames in training and validation data are from Kvasir v1 as Etis-Larib and CVC-Colon don't contain colorectal non-polyp frames.

were found unclear in the patient. The distal colon was the primary location of related adenomas, which were linked to advanced disease and numerous adenomas. According to the physician's opinion, it may serve as a possible indicator of the development of distant colorectal adenomas into cancer. The third video is about a 56-year-old woman who had irregular rectal bleeding for the past four months. A colonoscopy revealed a big lump with congested internal hemorrhoids. The fourth video is another clip from the first video of the same patient. The fifth video is about 59-year-old woman who had tubulovillous adenoma which is ileocecal in origin valve. The sixth video shows an adenoma. No further information is provided on the website. Finally, the seventh video shows an ileocecal valve-emerging sessile tubulovillous adenoma found in a 59-year-old female who underwent a colonoscopy as part of her routine medical care.

3.4.2 Proposed Architecture

The proposed architecture called `Window bAsed Detection afTer Mixed Convolutions Polyp Identification (WADT-MCPI)' consisted of an input layer with hyper dimensions, hyper channels, and depth dimensions, a feature extracting module, and two fully connected layers for identification of CP and non-CP frames. It is followed by a window-based detection of the CP region during test set analysis. The feature extracting module consisted of concatenations of one conventional convolutional layer with sixteen filters of kernel size 3×3, one layer of fractionally strided convolutions with thirty-two filters of kernel size 3×3, and one layer of depthwise separable convolution with sixty-four filters of kernel size 3×3; each layer followed by a max pooling layer with a pool size of 2×2. The organization of the convolutional layers in the feature-extracting module was experimentally checked using the feature maps and evaluation metrics (Table 3.2). After that, the learned features were flattened to a 1D vector and were passed to two fully connected layers of 128 and 1 units respectively for class identification purposes. All the layers except the last layer were activated with the help of the non-linear activation function “ReLU” which accelerates the convergence of stochastic gradient by using less activation while minimizing the cost function of the stated problem [117], [118]. A sigmoid function was utilized in the last fully connected layer as it helps in mapping the learned features between 0 and 1. It is often considered an appropriate unit for mapping output probabilities in a binary classification problem [118], [119]. After the evaluation stage, the proposed architecture was subjected to a window-based detection using the mentioned Algorithm 1 to detect the presence of CP in the four test sets. Initially, a window size of 64×64×3 was set to slide over CP colonoscopic frames of size

128×128×3. An empty matrix of the maximum box was initialized along with the maximum prediction set to zero. As the window slid over the pixels of each frame, the window patch was cropped and normalized to perform a prediction. The coordinates of the bounding box were computed. It was then superimposed as a red line, a rectangular-shaped box to represent the CP region in the original frame.



Algorithm 1 Window Based Detection Algorithm

```
$max\_pred = 0.0$
$max\_box = []$
$cropped\_img = a$
for $win\_size \in window\_sizes$ : do
    for $top \in range(0, img.shape[0] - win\_size + 1, step)$ do
        for $left \in range(0, img.shape[1] - win\_size + 1, step)$ : do
            $box = (top, left, top + win\_size, left + win\_size)$
            $a = img[box[0] : box[2], box[1] : box[3], :]$
            $a = a * 1./255$
            $a = a.reshape(1, a.shape[0, 1, 2])$
            $pred = model.predict(a)$
            if $pred[0][0] > max\_pred$ : then
                $max\_pred = preds[0][0]$
                $max\_box = box$
            end if
        end for
    end for
end for
```

3.4.3 Experimental settings

The proposed architecture has been implemented using Python with TensorFlow in the back end on a local machine with 16GB RAM of specification AMD Ryzen 7 2700 Eight-core processor, with NVIDIA Ge-Force GTX 1050Ti and 4GB RAM graphics card. A hyper image dimension, batch size, epochs, channels, and mode of 128×128, 32, 50, 3, and RGB respectively were set. The adam optimizer with a learning rate of 0.001, loss as binary cross entropy was set. A random seed was initialized to obtain reproducible tensors in the appropriate code lines. A total of 1,850,945 parameters were found to be trainable. The above-mentioned hyper-parameters have been considered after extensive hyper-parameter tuning and analysis. It is discussed in Section 3.5.1. Some of the formula for the utilized evaluation metrics have been mentioned below:

Classification accuracy is the percentage of properly defined data from the complete set indicated by the TP and TN condition or the percentage ratio of correctly specified data to the full dataset, as shown in Eq. (3.1).

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \qquad (3.1)$$

Here, TP is number of true positives, TN is number of true negatives, FP is number of false positives, and FN is number of false negatives.

Precision is the uniformity of the measuring findings as shown in Eq. (3.2).

$$Precision = \frac{T_P}{T_P + F_P} \qquad (3.2)$$

Recall is the proportion of similar occurrences recovered as shown in Eq. (3.3).

$$Recall = \frac{T_P}{T_P + F_N} \qquad (3.3)$$

F-score is the weighted harmonic mean of precision and recall as shown in Eq. (3.4).

$$F - score = \frac{Precision \times Recall}{Precision + Recall} \qquad (3.4)$$

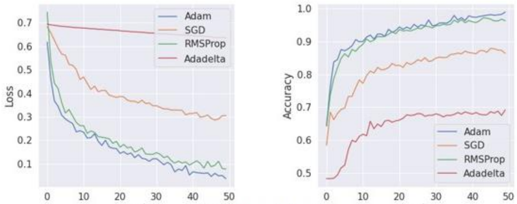
3.5 Results and Discussion

The present work proposes WADT-MCPI for automatic colorectal polyp diagnosis using colonoscopy frames. Several experiments were conducted to fine-tune the architecture and check its effect on the identification of the CPs in the validation set (presented in Section 3.5.1), conducted ablation experiments, feature maps, and CAM in Section 3.5.2. Section 3.5.3 discusses the effect of occlusion testing and four test sets on the proposed architecture. The comparative experimental results of the present work with vanilla Inception v3 architecture have been discussed in Section 3.5.4 followed by a comparative analysis with existing state-of-the-art works in Section 3.5.5.

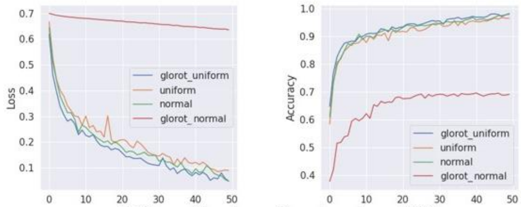
3.5.1 Hyper-parameter tuning

Four optimizers namely adam, stochastic gradient descent (SGD), RMSProp, and adadelta were considered while fine-tuning the proposed architecture optimizer. It can be noticed in Table 3.2 and Figure 3.3 that the Adam optimizer achieved the best results, followed closely by RMSprop, possibly due to its excellent robustness to the choice of hyper-parameters in comparison to the before-mentioned optimizers. Glorot uniform performed the best amongst the four kernel initializers due to its ability to achieve faster convergence in comparison to other initializers and detected 182 CP frames out of 215 frames. RGB colour space performed the best among the three-colour spaces namely gray and HSV. RGB colour space offers a linear combination of red, green, and blue spaces which is appropriate for the present pipeline as the chosen colonoscopy dataset, was originally captured with high resolution and intensity of colour spaces. While deciding the image dimension of the pipeline, two dimensions i.e., 128×128 and 64×64 were considered due to the highly variable image dimension of the chosen dataset. 128×128 dimensions were considered for further analysis as the architecture performed better in terms of average validation accuracy. A kernel size decides the dimensions of the filter matrix which slides over an image through various convolutional layers [120]. A 3×3 kernel size outperformed in comparison of the other kernel sizes depicting that smaller kernel sizes understand minute details and are able to achieve better receptive fields and further extract deeper features from the colonoscopic frames.

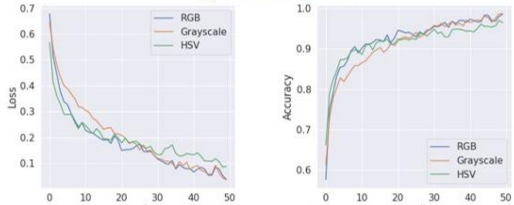
Hyper-parameter tuning of convolution layers is a computationally expensive task and required both theoretical and experiment-result-based understanding. So, in order to choose the best convolutional layer for the proposed feature extracting module, we tested four different convolutional layers namely the conventional convolution layer (referred to as conv2d in Keras), functionally strided convolution layer (referred to as conv2d transpose in Keras), depthwise layer (referred as depthwise conv2d in Keras), and depthwise separable (referred as separable conv2d in Keras) convolution layer. We built four separate modules from each of the before-mentioned convolution layers while keeping the same settings for the rest of the proposed work. The comparisons were made based on the run time and learnable elements of the layers, robustness towards polyp identification, and the developed feature maps.



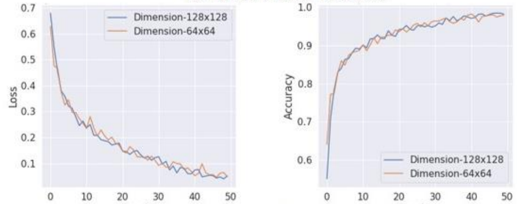
(a) With varied optimizers



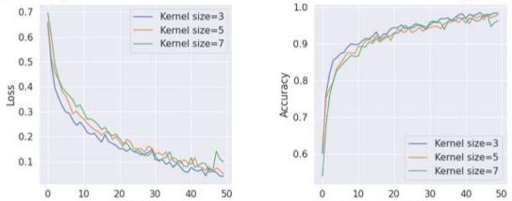
(b) With varied kernel initializers



(c) With varied color space



(d) With varied image dimension



(e) With varied kernel size

Figure 3.2 Reported results of the accuracy and loss for various varied hyper-parameters in the proposed feature extracting module.

Transpose convolution layer required the maximum no. of learnable elements (2,391,329) for the same settings followed by depthwise separable layer (1,609,068). While the least trainable parameters were obtained in the depthwise convolution layer (75,611) and were found computationally faster in comparison to other convolution layers. We also tested the dilated convolutional layer-based module with a dilation factor of 2. It produced similar results as conventional convolution layers in terms of training and validation accuracy but was found biased towards non-CP frames during test set analysis and hence not considered for the present study. It can be noticed from Table 3.2 that the separable convolution-based feature extracting module was able to identify all the CP frames but classified several non-CP frames as CP frames as well. Depthwise convolution layers were found biased towards non-CP frames and were not able to identify any CP frame in the validation set. The same can also be observed from the developed feature map of a precancerous CP frame which shows severe dysplasia in Figure 3.4. It is worth mentioning that appropriate hyper-parameter tuning and sequence decision of the proposed module, not only helped in extracting better feature abstractions in comparison with the conventional convolution-based module but also regulated the problem of vanishing gradient with increasing deeper filter and layers while reducing the spatial size of the input to reduce the no. of computations.

Table 3.2 Reported results of the accuracy and loss for various varied hyper-parameters in the proposed feature extracting module.

Parameter	**Average training accuracy**	**Average training loss**	**Average validation accuracy**	**Average validation loss**	**No. of polyps detected correctly (215)**
(a) Optimizers					
Adam	93.01%	0.16	91.66%	0.25	206
SGD	81.56%	0.39	82.95%	0.37	120
RMSProp	91.58%	0.20	88.62%	0.32	192
Adadelta	65.80%	0.66	62.74%	0.67	103
(b) Kernel initializers					
Glorot_uniform	92.77%	0.16	90.93%	0.25	182
Uniform	91.02%	0.21	88.54%	0.31	141
Normal	91.82%	0.18	90.74%	0.27	156
Glorot_normal	65.17%	0.66	64.72%	0.67	81
(c) Colour space					
RGB	92.25%	0.17	90.60%	0.26	206
Grayscale	91.08%	0.20	84.84%	0.51	103
HSV	91.30%	0.19	51.27%	6.05	98
(d) Image dimension					
128 × 128	92.30%	0.17	95.54%	0.29	202
64 × 64	92.62%	0.17	89.74%	0.22	201
(e) Kernel size					
3	92.75%	0.16	89.7%	0.27	183
5	91.51%	0.19	89.45%	0.28	144
7	91.06%	0.20	89.34%	0.30	84
(f) Convolution layers					
Separable Conv2D	83.13%	0.36	83.74%	0.37	215*
Depth wise Conv2D	78.78%	0.43	80.14%	0.44	0*
Transpose Conv2D	86.28%	0.30	87.43%	0.33	101
Conventional Conv2D	87.84%	0.28	89.85%	0.26	182

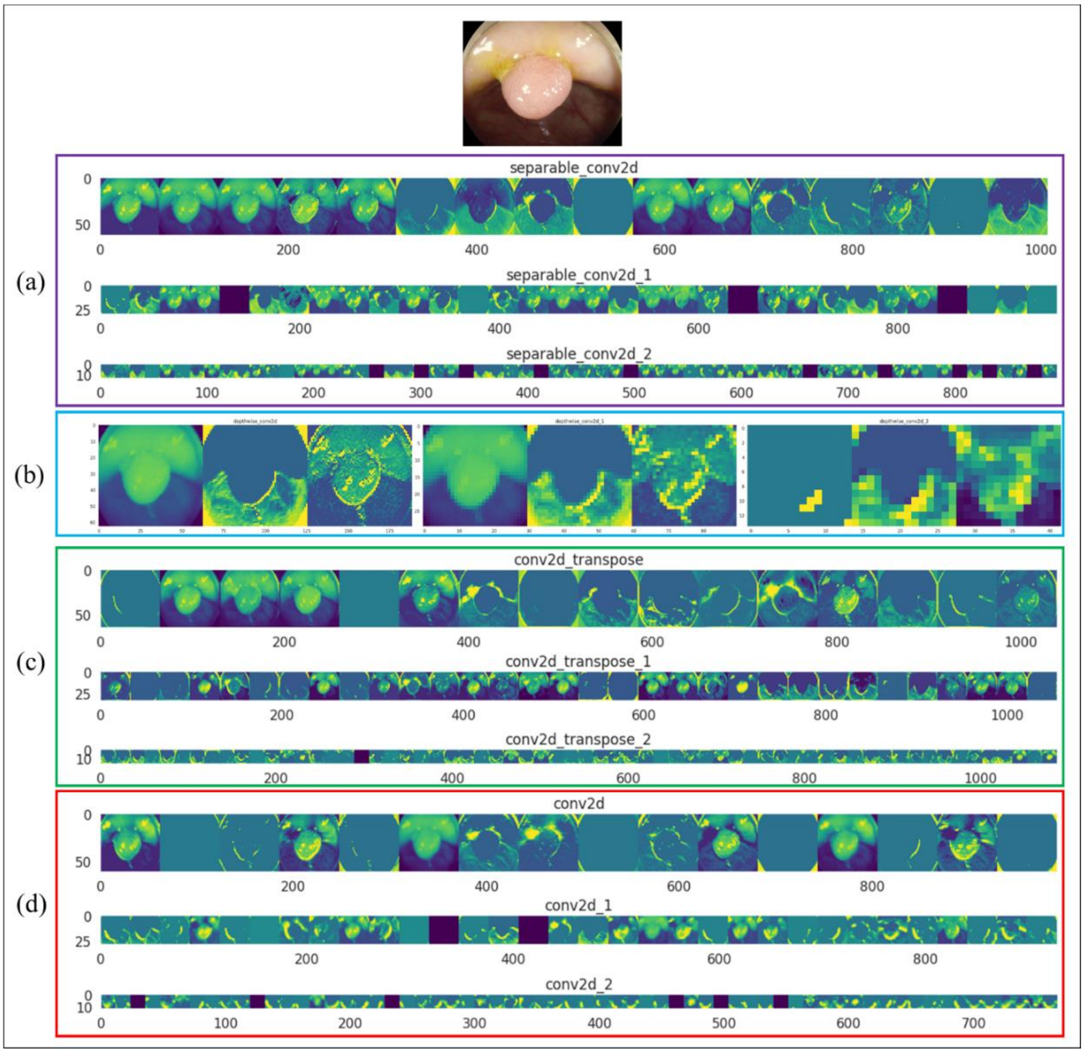


Figure 3.3 Achieved feature maps of the different convolution layers. (a) Depthwise separable layer, (b) Depthwise layer, (c) Transpose layer, and (d) Conventional convolution layer.

3.5.2 Feature mapping, class activation mapping and ablation experiments

A feature map represents the output of applying a filter to an input image. It helps in visualizing and explaining the complex, internal representations detected by the convolution filter layer for that input image [121]. Several feature maps were developed while fine-tuning the proposed feature-extracting module. Based on the feature map observations and achieved results from the validation set, the present work developed the sequence of convolution layers in the proposed feature extracting module. First, we discuss the feature maps in Figure 3.4. It can be noticed that both depthwise separable and conventional convolution layer has the deadest filters represented as completely dark blue followed by the transpose layered module. However, both layered modules have several activations on edges, and boundaries within the image. Some of the filters were perfectly able to detect the shape of the precancerous CP frame.

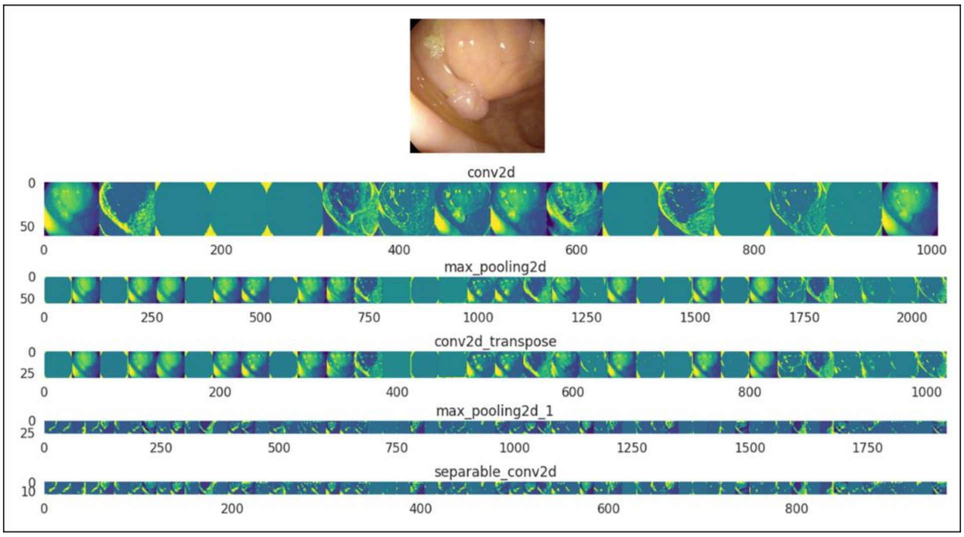


Figure 3.4 Achieved feature maps of the proposed feature extracting module.

Depthwise separable convolutions coming from the Inception family are inspired by the estimation of matrices through the merging of two convolution operations namely the spatial convolution and point-wise convolution operation [120], [121]. They help in learning higher levels of feature abstraction with increasing depth dimension [117], [120]. Functionally strided convolutions also referred to as de-convolutions or transposed convolutions help in projecting the obtained feature maps to a higher dimensional space by swapping the forward and backward passes of a convolution process [120]. It doesn't necessarily obtain the opposite of a conventional convolutional layer but helps in maintaining connectivity of the previous feature map with the next by maintaining the same matrix shape [120]. The same can be observed from the obtained feature maps in Figure 3.4 and 5. Figure 3.5 depicts the feature maps obtained after fine-tuning the proposed feature-extracting module. It can clearly be noticed that the no. of dead filters has decreased because of the chosen sequence of the proposed feature extracting module wherein even the reduced input image after the second pooling stage was able to extract deeper features such as colour, depth, shape, and boundaries of the CP frame.

CAM is another method to develop trust and transparency of the CNN layers for the health officials by providing a weighted map of the input image which depicts the important regions utilized while predicting new images for a certain layer without explicitly declaring the bounding boxes [122]. Grad CAM is a generalization of CAMs and uses the classification score gradient to determine which parts of the image are most crucial for classification (here, identification of CP) [123]. Figure 3.6 shows some of the generated CAMs from the training, testing, and validation set of the present work. The red colour depicts the parts of the proposed architecture exhibiting high attention and are a CP frame during the prediction stage. The blue region emphasizes less attention and firing of the activation function at the region. The proposed architectural layers perfectly emphasized the CP region for the first three CP frames in Figure 3.6 while the fourth frame's CP region was not understood well. A discussion on the missed polyps has been done in the next sub-section.

The ablation experiments were conducted in a systematic manner, wherein the layers in the feature-extracting module were commented on one by one and checked for all evaluation metrics during the architecture fitting stage. An average decrease of 5-10% was observed in the validation accuracy with respect to baseline accuracy of 92.29% when any of the layers were removed. To analyze the effect of hyper-parameters mentioned in Section 4.1 and vanilla Inception v3 architecture results in up to 50 epochs have been discussed.

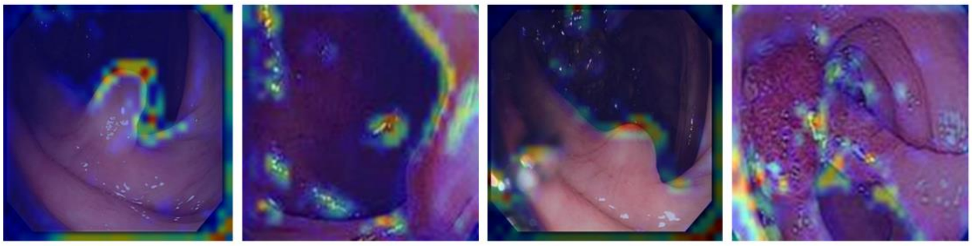


Figure 3.5 Class activation maps.

3.5.3 Test set analysis and occlusion testing

Test set analysis helped us to evaluate the future, real-time performance of the proposed architecture with unaltered, un-seen colonoscopic data during the training and validation stage. Table 3.3 shows the no. of correctly predicted CP frames from all the test cases. First, we discuss the results achieved for test 0 consisting of the developed `Gastrointestinal atlas-Colon Polyp' test data. It consists of sequential frames of seven patient videos extracted from an open-source medical website. There are several qualities to the developed test set i.e., a) the presence of de-identified (non-sequential) frames for future application of AI techniques; b) the presence of sequential positive and negative CP video frames for sequential frame analysis for both medical and AI analysis; c) Variety of CPs in terms of size, type, and its location; and d) Same image dimension (128×128) with high-resolution and processed data with no patient information in the released frames. We also release the full videos in a separate folder for future medical video analysis.

Figure 3.7 shows some of the missed CP windows (thirty-four) for test 1 which consisted of randomly selected, positive frames from polyp-gen data. Upon close observation, it can be noticed that most CPs are flat, sessile serrated in nature [83] or polyp smaller than 1 mm, present on the folds of the mucosal layer or hidden due to reflection of a camera at the time of colonoscopic recordings. All such frames may be difficult to detect due to indifference in colour with the colon surface. There may be several reasons behind this missed rate i.e., a) polyp-gen data consists of mixed-sized polyps while the training and validation set contains less variation of small-sized polyps. b) considered window size during the detection stage. Test 2 consisted of mixed frames with variable settings of colonoscopy, resolution, origin, types, and size of polyps. Figure 3.8 (c) and (d) depict an example of correctly detected and identified CP sequential frames for videos 6 and 7. It can be noticed that the proposed architecture was able to identify and detect CPs in varying environments. Figure 3.8 (a) shows the correctly detected CPs from test 2 wherein it can be observed that the proposed architecture was able to detect multiple minute and large-sized CPs. Figure 3.8 (b) shows the CP frames which were common in tests 0 and 2 and were repeatedly missed in both the test set analysis. The fourth CP frame in Figure 3.6 verifies the reason behind this wherein in the fourth CAM, the higher activation can be observed around the boundary of the polyp region. The introduction of a variety of CP and non-CP frames along with data augmentation methods such as zooming out to reduce the size of the CPs may correct this missing rate.

To the best of our knowledge, occlusion testing has not been carried out in this field.

It refers to hiding or blocking or closing a certain portion of a frame that may be otherwise important for prediction. Such an analysis also helps in interpreting predictions done by AI networks. The results of the test set analysis of Test 3 which consisted of the developed occlusive frames shown in Figure 3.1 have been prepared based on the discussion and opinions of an experienced gastroenterologist. We discuss two interesting cases depicted in Figure 3.8 (e) where the first frame contains a benign polyp in its progressive stage and the second frame is an adenoma with high-grade dysplasia. The proposed architecture predicted both frames as non-CP frames. However, in the opinion of the experienced gastroenterologist, the first frame may be considered a non-CP frame but the stalk of the adenoma in the second frame is still visible and may not be overruled as a non-CP frame.

Table 3.3 Results of the test set analysis.

Test Set No.	Total no. of polyps	Correctly predicted polyps
Test 0	150	139
V1. Polipo MSD z6	15	11
V2. Pediculado 3	38	35
V3. Polipo MSD z2	23	21
V4. Polypileocecalval ve1	22	21
V5. Polypvv 5	24	23
V6. Rectalcarpet1	16	16
V7. Pediculado 5	12	12
Test 1	256	222
Test 2	103	68
Test 3	10*	7*

Similar to this, the present work also developed four frames to analyze the AI prediction of frames after the complete resection of a polyp. A recent study suggested that AI networks misidentified such frames and produced ambiguous results [89]. To check this, we completely patched four CP frames from the training set with their adjacent colon surface. The proposed architecture identified three frames as non-CP frames out of the four frames. However, the generalized conclusion cannot be made and remains a future prospect for the present work. Studies have proposed different positive and negative augmentations in this field to increase the quantum and variety of data [124], [125], [126]. Through the present analysis, it can be observed that such augmentations should be introduced with care as they may hide the polyp region and appear `normal', confusing the AI network as a non-CP frame leading to ambiguous results and developing mistrust among health professionals.

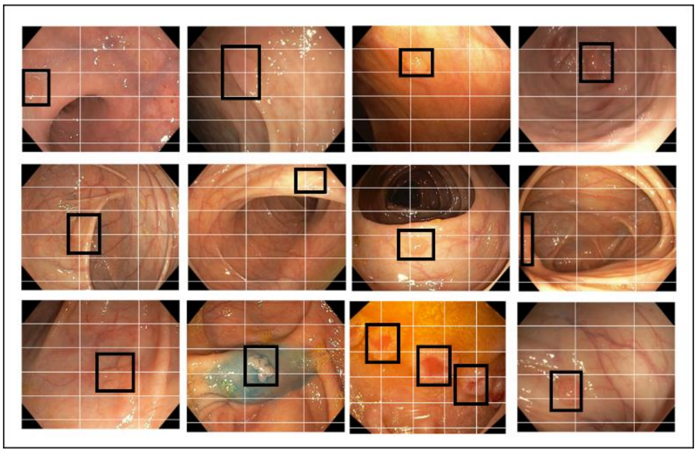


Figure 3.6 Missed polyp frames during test set analysis from Polyp-gen dataset.

3.5.4 Comparison with vanilla Inception v3 architecture

The present work also compared the achieved results with the famous Inception v3 architecture as it has proven to achieve efficient prediction for the Image Net database with a lower error rate and is found computationally less expensive in comparison to several transfer learning architectures [39], [127], [128], [129], [130]. The term vanilla refers to the development of standard Inception v3 architecture without fine-tuning. An accuracy, precision, recall, specificity, and F1-score up to 48.67%, 50.77%, 30.70%, 6-8.00%, and 38.26% respectively was achieved on the colonoscopy dataset for Inception v3 architecture. The present work has achieved accuracy, precision, recall, specificity, and F1-score up to 92.53%, 94.23%, 91.16%, 94%, and 92.67% respectively. Clearly, the proposed architecture outperformed the vanilla Inception v3 for all evaluation metrics. Figure 9 depicts the achieved loss and area-under-curve (AUC) score graph of the Inception v3 architecture. It was observed that the proposed architecture not only computationally ran faster but also achieved a steadily decreasing loss graph for both the training and validation set for increasing no. of epochs. Further, the proposed architecture required less no. of trainable parameters (1,850,945) as opposed to the vanilla Inception v3 architecture (21,819,170).

3.5.5 Comparison with existing state-of-the-art works

A comparative analysis of the proposed architecture has been done with the recent state-of-the-art works from 2020-2022 in automatic polyp diagnosis using colonoscopy frames. Comparative parameters like the similarity of the dataset, CNN architecture and estimated explainability methods, evaluation metrics, and computational resources have been considered in Table 3.4.

Patel *et al.* [131] performed a comparative analysis between VGG19, ResNet, DenseNet, SENet, MnasNet, on four different, private, and publicly available datasets for automatic polyp classification. In comparison with their best-reported results with the proposed architecture, the present work has achieved higher precision, AUC score, and overall accuracy while identifying CP frames. In addition, the present work has included several explainability methods and evaluation metrics to test the robustness of the proposed architecture.

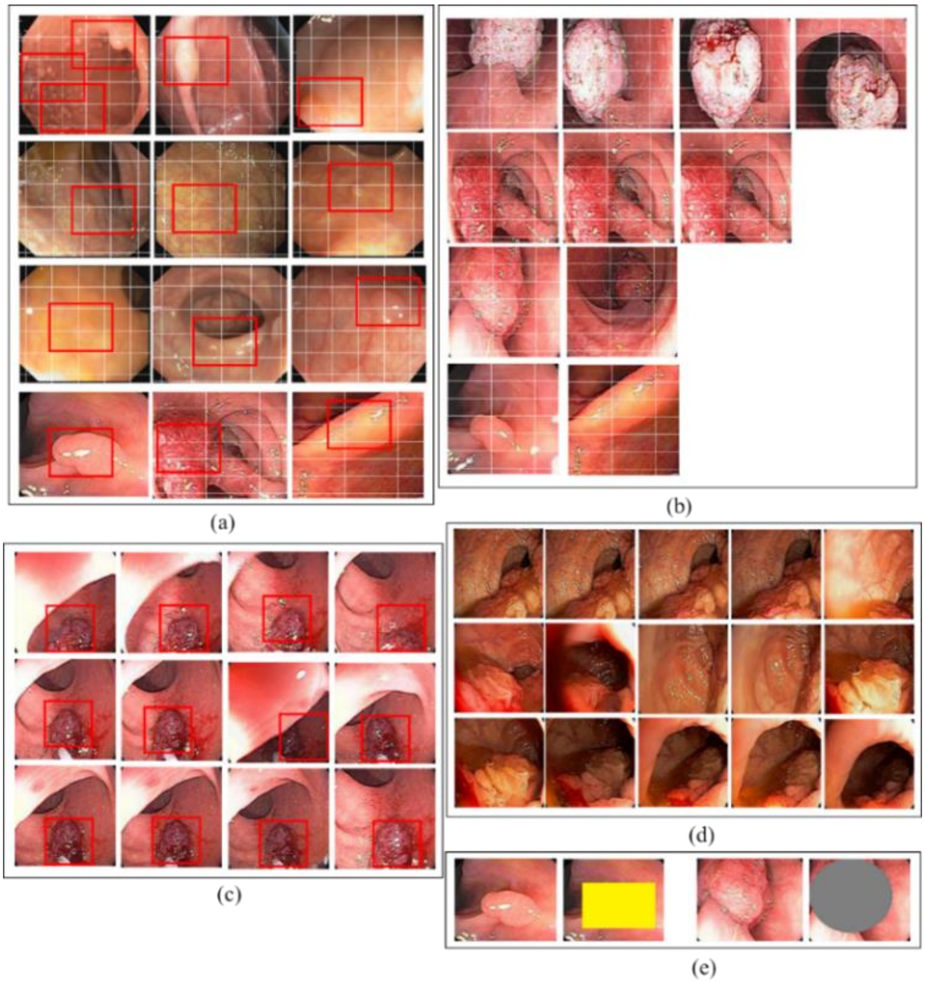
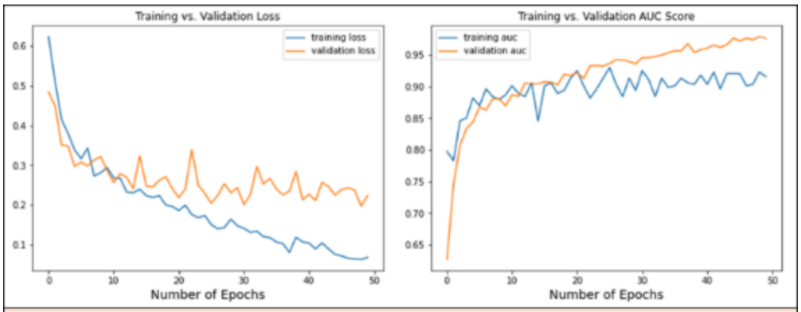


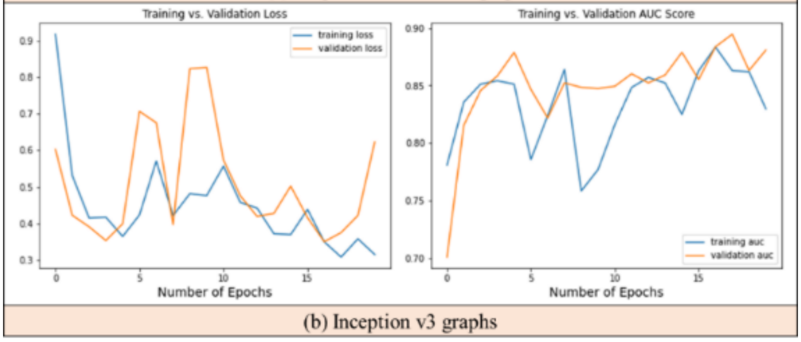
Figure 3.7 Colorectal polyp (CP) frame examples of the results achieved during test set analysis. (a) Correctly detected CP frames. (b) Repeatedly missed CP frames. (c) and (d) Correctly detected and

identified frames. (e) Occlusive frames.

Jheng *et al.* [132] proposed a CNN-based algorithm to identify colonic diseases and normal anatomical landmarks using privately collected colonoscopic frames. They have achieved a slightly higher recall, specificity, and overall accuracy in comparison to the proposed architecture. However, it is worth mentioning that their proposed CNN-based algorithm with VGG16 backbone (GUTAID) has not been validated on any open-source dataset nor tested on un-altered, varied colonoscopy sequential and non-sequential frames which were not a part of training and validation set. On the contrary, the present work has focused on performing an extensive test set analysis on varied CP types, sizes, shapes, and sources while also checking the effect of each layer with respect to CP identification in colonoscopy frames along with systematic ablation studies. This further shows the generalizability of the proposed architecture.



(a) Proposed architecture graphs



(b) Inception v3 graphs

Figure 3.8 Achieved loss and AUC score graphs for the proposed architecture and vanilla Inception v3.

Jia *et al.* [133] proposed a PLPNet model for automatic polyp recognition and segmentation using publicly available CVC-ColonDB, CVC-ClinicDB, and GIANA 2017 databases. Their pipeline executed each frame within 380 ms on a single NVIDIA GeForce GTX TITAN Xp. The proposed architecture was able to execute each frame in comparatively less time (300 ms) on the eight-core processor, NVIDIA GeForce

GTX 1050Ti GPU. The achieved recall is slightly less (0.94%) than [133], possibly due to the consideration of different datasets with less no. of non-CP frames. Ellahyani *et al.* [134] proposed a fine-tuned polyp detection pipeline with a hybrid mixture of VGG16 and mobile net using ETIS-LaribPolypDB, kvasir-seg, and CVC-ClinicDB database. They have achieved a higher AUC score value possibly due to the imbalance in CP and non-CP frames in their chosen dataset. The present work has outperformed in terms of all the other comparative parameters such as precision, recall, F1 score, and explainability methods. It can be noticed from Table 3.4 that the proposed architecture has performed better in comparison to the work done by Rahim *et al.* [114] for all the comparative parameters.

Table 3.4 Comparison of the proposed pipeline with existing state-of-the-art works. NR = not reported.

Parameters	[131]	[132]	[133]	[134]	[114]	**Proposed**
Precision (%)	78.09	NR	84.80	91.00	94.44	94.23
Recall (%)	NR	89.80	92.10	89.00	82.92	91.16
Specificity (%)	NR	96.80	NR	NR	NR	94.00
F1 Score (%)	NR	NR	88.30	90.00	88.30	92.67
AUC Score	76.40	NR	NR	95.20	90.42	91.75
Overall accuracy (%)	75.70	93.30	NR	NR	NR	92.53
Execution time per epoch (milli seconds)	NR	NR	381 ms	NR	600 ms	300 ms
Ablation study	No	No	No	No	No	Yes
CAM graphs	No	Yes	Yes	No	Yes	Yes
Occlusion testing	No	No	No	No	No	Yes
Separate test set analysis	Yes	No	No	No	No	Yes
Feature mapping	No	No	No	No	No	Yes
Hyper-parameter tuning	No	Yes	Yes	Yes	Yes	Yes

3.6 Conclusion and Future Scope

In this work, an explainable, end-to-end WADT-MCPI architecture for automatic colorectal polyp diagnosis has been proposed using colonoscopy CP and non-CP frames. The proposed architecture consists of a novel, fine-tuned feature-extracting module, followed by CP and non-CP frame identification and a window-based CP detection system. The work has achieved an overall accuracy, precision, recall, specificity, F1 score, and AUC score up to 94.23%, 91.16%, 94.00%, 92.67%, 91.75%, and 92.53% respectively. A new test set has also been developed and released for research purposes. Explainable and evaluation methods like class activation mapping, feature mapping, occlusion testing, hyper-parameter tuning ablation experiments, and separate, sequential, and non-sequential frame-based test set analysis have been used to show the efficacy of the proposed architecture. Future works will focus on improving the evaluation metrics, and inclusion of more colonoscopy data with different types of occlusion effect.

CHAPTER 4

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE KOREA CANADA FOR CLEANLINESS ASSESSMENT IN ENDOSCOPY

4.1 Introduction

Conventional radiologic and endoscopic techniques have difficulty penetrating the small bowel. VCE a non-invasive, non-anesthetic diagnostic technique is utilized for viewing inside the small bowel due to its minuscule size [14]. It is a relatively a newer technique in comparison to other endoscopic techniques which was developed in 1981 and received clinical approval in 2001 [14]. It is commonly used in obscure GI bleed, iron deficiency anemia, Crohn’s disease, celiac disease, familial syndromes etc. However, there are limitations to VCE including its cost, longer reading time, high miss-rate of lesion detection, and no therapeutic applications [20], [21], [23], [25], [66], [135], [136].

The high miss-rate of lesion detection is due to several limitations. One of the major limitations behind this is the lack of adequate bowel preparation, and its objective scoring system [23], [25]. This limitation is common in other endoscopic procedures as well. However, no strict guidelines have been proposed by the inventors/ companies of VCE to assess the ‘adequacy’ of the bowel. An adequate bowel preparation and its objective scoring system is essential for obtaining and analyzing a good quality VCE video for lesion detection [23].

There is currently a lack of a valid, objective, fast, repeatable, and dependable scoring system to evaluate the bowel’s sufficiency in VCE [23], [137]. A number of scores, including the Viazis score (2004) [138], Brotz score (2009) [139], Park’s score (2010) , and KODA score (2020) [140], have been proposed to evaluate the small bowel preparation. These scores are state-of-the-art, laborious, and rely on the observer’s judgment to be scored. They have not been standardized for picture or video analysis and are not frequently utilized in conventional clinical practice.

4.2 Related Works

We discuss Viazis score (2004) [138], Brotz score (2009) [139], Park's score (2010), and KODA score (2020) [140] which have been proposed to evaluate the small bowel preparation.

Viazis *et al.* [138]conducted a prospective, randomized, controlled trial to investigate the impact of bowel preparation on the diagnostic yield of CE. They enrolled patients and randomly assigned them to either receive bowel preparation or not before undergoing CE. The researchers utilized a standardized bowel preparation protocol, which included a combination of laxatives and clear liquid diet. Following the procedure, they employed a scoring system to assess the efficacy of bowel preparation based on the cleanliness of the small bowel mucosa based on adequacy (adequate v/s non-adequate).

The results demonstrated a significant increase in the diagnostic yield of CE among patients who underwent bowel preparation compared to those who didn't. Specifically, the group that received bowel preparation exhibited a higher detection rate of small bowel lesions. This finding emphasized the importance of proper bowel preparation in optimizing the diagnostic accuracy of CE for identifying small bowel disorders. Moreover, the study underscores the effectiveness of the standardized bowel preparation protocol utilized, as indicated by the cleanliness scores of the small bowel mucosa. These findings highlight the necessity of incorporating bowel preparation as a routine component of CE procedures to enhance diagnostic outcomes and improve patient care.

Brotz *et al.* [139] aimed to validate three grading systems for assessing small bowel cleansing in CE. The study encompassed a quantitative index, a qualitative evaluation, and an overall adequacy assessment. Patients undergoing CE were prepared with a standard bowel cleansing regimen consisting of a polyethylene glycol-based solution and a clear liquid diet. The researchers employed three grading systems to evaluate the quality of small-bowel cleansing: a quantitative index based on CE images, a qualitative assessment by expert reviewers, and an overall adequacy assessment integrating both quantitative and qualitative aspects. The findings of the study demonstrated the effectiveness of the standardized bowel preparation protocol in achieving adequate small-bowel cleansing. Moreover, the validation of the three grading systems revealed their utility in accurately assessing the cleanliness of the small bowel during CE procedures. The study underscores the importance of utilizing standardized grading systems to evaluate bowel preparation efficacy in CE, providing clinicians with valuable tools to optimize diagnostic accuracy and patient outcomes.

Park *et al.* [141] presented a novel cleansing score system for CE aimed at evaluating the effectiveness of bowel preparation. Prior to CE, patients underwent preparation involving a polyethylene glycol-based solution and a clear liquid diet, ensuring optimal visualization of the small bowel mucosa. The novel cleansing score system introduced by the authors involved assessing four key aspects of bowel cleanliness: fluid accumulation, residual bubbles, small bowel visualization, and presence of debris. Each criterion was assigned a score ranging from 0 to 2, with higher scores indicating better cleansing. The total score ranged from 0 to 8, with higher scores reflecting superior bowel cleanliness. The study found that the novel cleansing score system provided a comprehensive and objective means of evaluating bowel preparation efficacy for CE. Moreover, the system demonstrated good interobserver agreement among the reviewers, enhancing its reliability in clinical practice.

KODA score is an extension of Park score with variation in the range of the questions asked in KODA [140], [141]. It is the latest manual scoring system for assessing the small bowel preparation quality in VCE. It has been clinically validated on 1,233 still images obtained from twenty-five capsule videos with the help of twenty health experts. The still images were captured on a five-minute interval from the first picture of the duodenum to the first picture of the cecum during each VCE procedure i.e., in a four-hour small bowel transit time, forty-eight images were captured and then scored. The final scoring is based on the score achieved from two sub-scores namely the percentage of visualized mucosa and degree of obstruction in each VCE frame divided by the total number of frames captured in one VCE procedure. A standardized training module has been released by the inventors of the KODA score to acquaint the score system to the health experts in this field. The training module is currently state-of-the-art and has been utilized to assess the outcome of one randomized controlled study in VCE [142].

Automatic computer-operated assessments have also been introduced in this field. The studies have focused on comparing the red over green pixel ratio (R/G), R/G ratio, abundance of bubbles, brightness with the help of machine learning, checking adequacy and in-adequacy using neural networks, automatic score classification of visible mucosa, dirty and clean patches, and colour bar differentiation. These studies may prove to be more reliable, quick-to-assess, and provide independence from the intra- and inter- observer variability.

4.3 Problem Statement

Currently, no computer-operated assessment study has been done to automate an existing score for cleanliness assessment in VCE. The primary objective of the present study was to develop an automated approach to automatically assess the cleanliness of VCE as per the latest scoring system using AI models in real-time. The secondary objective of the present study was:

- To develop a simple and user-friendly application for gastroenterologists to score the VCE frames as per the latest scoring system i.e., KODA. The application is called as AI-KODA score. It is fully auto- mated and works in real-time on an Android phone. It also helps in development of a multi-labelled im- age dataset.
- To generate a high-quality, multi-labelled image dataset which is medically validated and can be utilized for development of AI models for computer- operated assessments in this field.
- To conduct a comprehensive evaluation, interpretation, benchmarking of the generated dataset using famous AI algorithms.

In this chapter, the first objective will be covered wherein we will focus on the development of AI-KODA, its study design followed to collect the scores through AI-KODA, determination of inter-rater and intra-rater reliability of KODA score among three readers for prospective AI applications. The AI applications will be discussed in the next chapter.

4.4 Methodology

In this, we will discuss the preparation of the capsule videos, followed study design, and statistical analysis setup in Section 4.4.1, 4.4.2, and 4.4.3 respectively.

4.4.1 Preparation of capsule videos

Twenty-eight patient capsule videos performed at a tertiary care academic institute were used in this study. All small bowels capsule endoscopic procedures were performed using Pillcam SB3 capsule (Given imaging) and read on RAPID viewer. All the patients had received a bowel preparation as per the standard guidelines. In each VCE video, the abnormality was diagnosed and marked, and anatomical land markings were present. The VCE frames were selected at an interval of five-minutes as per KODA from each VCE video. Random frames were also selected which consisted of abnormality and anatomical landmarks by experienced gastroenterologists irrespective of the obstruction or amount of visibility present in the VCE frame during the small bowel transit. In total, 1539 sequential-frames and 634 random frames were selected in twenty-eight videos. Both the frames were exported in high-quality jpg format, cropped and resized from 576 × 576 to 320 × 320. Then both the frames were uploaded into the backend (Microsoft Azure) of AI-KODA Score testing module.

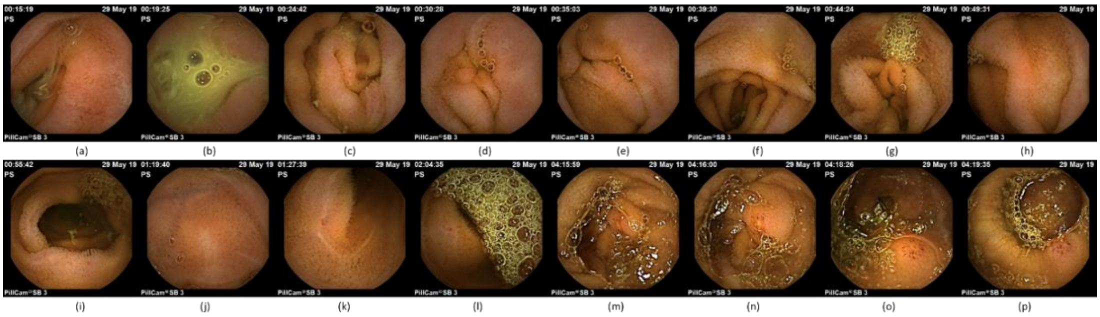


Figure 4.1 Example of CE frames selected over (a) – (h) five-minute interval and (i) – (p) random interval in one capsule video using the RAPID viewer software.

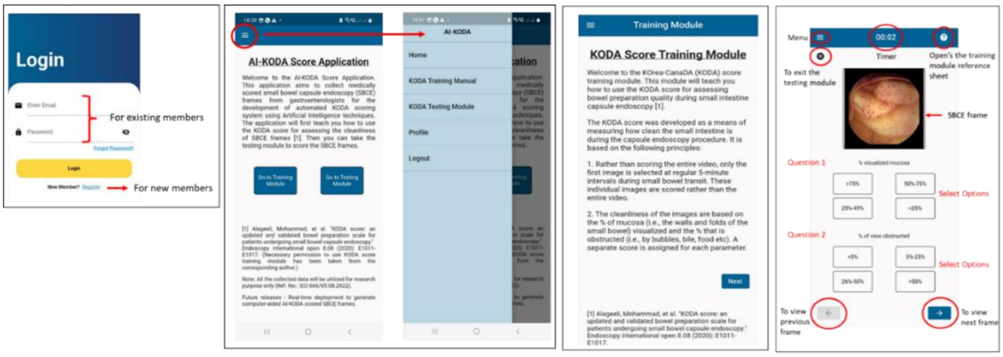
All the videos and extracted image frames were anonymized and all the information related to the patient was removed. Figure 4.1 depicts the selected frames. The study was done according to Helsinki declarations and was approved by the institute ethics committee (Ref. No.: IEC-666/05.08.2022) and a waiver of consent was granted.

4.4.2 Study design

Three gastroenterology fellows who had been trained in reading VCE shared a manual and a link to download the AI-KODA score application in their mobile phone. The application consisted of a secure login system, profile setup, and two modules namely training module and testing module. Figure 4.2 depicts the AI-KODA score application. The training module was taken from the original KODA score after necessary permission from the authors [140]. The testing module consisted of 2173 frames which were selected from the twenty-eight patient capsule videos. Two questions i.e., percentage of mucosa visualized, and degree of obstruction were displayed on each VCE frame. In the first question, the four options were $> 75\%$ (representing VMscore1), $50\% - 75\%$ (representing VMscore2), $25\% - 49\%$ (representing VMscore3), and $< 25\%$ (representing VMscore4). In the second question, the four options were $< 5\%$ (representing OVscore1), $5\% - 25\%$ (representing OVscore2), $26\% - 50\%$ (representing OVscore3) and $> 50\%$ (representing OVscore4). The option selected for each of the question, their timestamp and email ID were saved in real-time in the application's backend. In the back end, for each selection in any of the two questions, a numeric '1' was assigned. The rest of the non-selected options were assigned a numeric '0'. In this manner, each VCE frame was assigned two labels (VM and OV sub-scores) out of the total eight labels. This assignment was done by the inspiration of one-hot encoding method. It is a method used to convert categorical values to binary value of '0' or '1'. Figure 4.3 represents the front-end and backend flow of the developed application. Features of training module includes:

- User-friendly manual to learn and understand the KODA Score.
- Four examples and their correct answers to get familiar with the KODA Score.
- Features of testing module includes:
- User-friendly Testing Module to score the VCE frames on the screen.
- The KODA Score reference sheet can be viewed from the top question mark (?) button for any ambiguity/help to score the frame.
- A timer to note the time taken to score one frame.
- Forward button to view the next VCE frame.

- Backward button to view the previous VCE frame.
- A pop-up in case the user has not selected ANY option from the two questions.
- Exit button to exit the Testing Module.



AI-KODA Score Application

Welcome to the AI-KODA Score Application. This application aims to collect medically scored small bowel capsule endoscopy (SBCE) frames from gastroenterologists for the development of automated KODA scoring system using Artificial Intelligence techniques. The application will first teach you how to use the KODA score for assessing the cleanliness of SBCE frames [1]. Then you can take the testing module to score the SBCE frames.

[1] Alageeli, Mohammad, et al. "KODA score: an updated and validated bowel preparation scale for patients undergoing small bowel capsule endoscopy." Endoscopy international open 8.08 (2020): E1011-E1017. (Necessary permission to use KODA score training module has been taken from the corresponding author.)

Note: All the collected data will be utilized for research purpose only (Ref. No.: IEC-666/05.08.2022).

Future releases - Real-time deployment to generate computer-aided AI-KODA scored SBCE frames.

KODA Score Training Module

Welcome to the KOrea-CanaDA (KODA) score training module. This module will teach you how to use the KODA score for assessing bowel preparation quality during small intestine capsule endoscopy [1].

The KODA score was developed as a means of measuring how clean the small intestine is during the capsule endoscopy procedure. It is based on the following principles:

1. Rather than scoring the entire video, only the first image is selected at regular 5-minute intervals during small bowel transit. These individual images are scored rather than the entire video.

2. The cleanliness of the images are based on the % of mucosa (i.e., the walls and folds of the small bowel) visualized and the % that is obstructed (i.e., by bubbles, bile, food etc). A separate score is assigned for each parameter.

[1] Alageeli, Mohammad, et al. "KODA score: an updated and validated bowel preparation scale for patients undergoing small bowel capsule endoscopy." Endoscopy international open 8.08 (2020): E1011-E1017.

(a) (b) (c) (d)

Figure 4.2 Snapshot of the of the developed application. (a) Login page, (b) home page, (c) training module, and (d) testing module with its different operations.

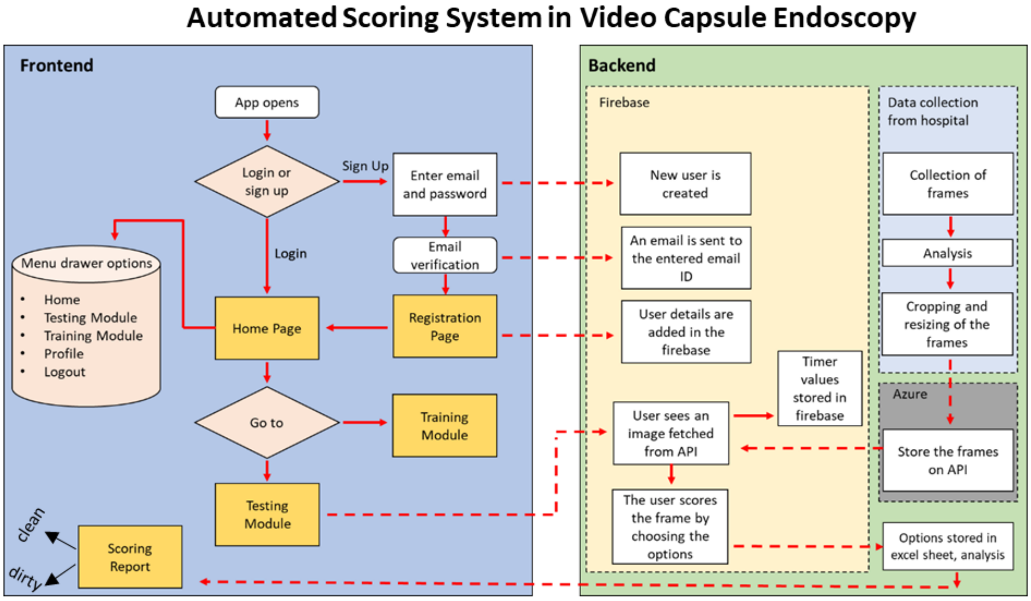
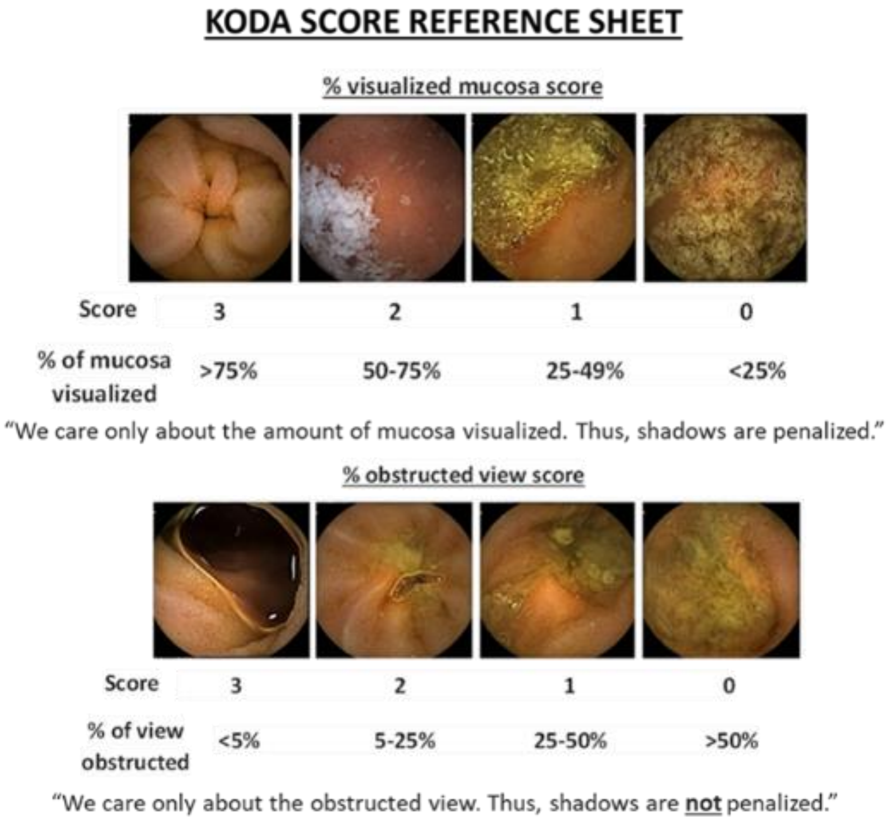

Automated Scoring System in Video Capsule Endoscopy
Frontend
App opens
Login or sign up
Sign Up
Enter email and password
Login
Email verification
Menu drawer options
Home
Testing Module
Training Module
Profile
Logout
Home Page
Registration Page
Go to
Training Module
Testing Module
Scoring Report
clean
dirty
Backend
Firebase
New user is created
An email is sent to the entered email ID
User details are added in the firebase
Timer values stored in firebase
User sees an image fetched from API
The user scores the frame by choosing the options
Data collection from hospital
Collection of frames
Analysis
Cropping and resizing of the frames
Azure
Store the frames on API
Options stored in excel sheet, analysis

Figure 4.3 Frontend and Backend Data Flow Storage System of AI-KODA Score Application.

The fellows were allowed to toggle between the frames and could change their answer by clicking on the backward and forward button on the testing module. Both entries were saved and analyzed for inter-rater and intra-rater reliability estimates in IBM SPSS Statistics 20. The study was conducted twice with a gap of four weeks in between them.

All the three fellows were asked to sign up for the application, set up a profile and complete the training module before starting the testing module both the times. They were also given an option to access the training module and the KODA reference sheet at any point of time during the scoring stage. Figure 4.4 represents the KODA reference sheet. All the fellows' independently rated 2173 frames as per KODA score and were unaware of clinical information related to the frames. Upon completion of the scoring, the fellows were given a feedback form to assess the efficacy of the developed application, its interface and automation with the help of Likert scale. They were asked to provide suggestions for its further development as per real-time clinical settings.



KODA SCORE REFERENCE SHEET

% visualized mucosa score

Score	3	2	1	0
% of mucosa visualized	>75%	50-75%	25-49%	<25%

"We care only about the amount of mucosa visualized. Thus, shadows are penalized."

% obstructed view score

Score	3	2	1	0
% of view obstructed	<5%	5-25%	25-50%	>50%

"We care only about the obstructed view. Thus, shadows are **not** penalized."

Figure 4.4 KODA score reference sheet.

4.4.3 Statistical Analysis Setup

Intra-class correlation coefficients (ICCs) are often utilized as a quantitative estimate to check the aspect of reliability. Eq. (4.1) represents the way of calculating ICCs.

$$\mathrm{ICC} = \frac{variance\ of\ interest}{total\ variance} = \frac{variance\ of\ interest}{variance\ of\ interest + unwanted\ variance} \qquad (4.1)$$

So, the ICCs were estimated using two-way random effect model described in [143] with 95 % confidence interval (CI) and measure of consistency using IBM SPSS Statistics 20. The strength of the achieved estimates was interpreted as poor if the ICC value was less than 0.5, moderate if the ICC value was between 0.5 and 0.75, good if the ICC value was between 0.75 and 0.9 and excellent for ICC value above 0.9 [143]. The sample size of twenty-eight videos was calculated as per ICC hypothesis testing with a minimum acceptable reliability, expected reliability, significance level, number of raters, and repetitions per subject as 0.8, 0.9, 0.05, 3, and 2, respectively [143]. The Likert scale was interpreted as excellent if the rating was 5, very good as 4, good as 3, fair as 2, and poor if the rating was assigned a value of 1 [144].

4.5 Results

In this, we discuss the demographic and statistical details of the patients, and AI-KODA score application feedback and efficacy and reliability estimate in Section 4.5.1, 4.5.2, and 4.5.3 respectively.

4.5.1 Demographic and statistical details of the patients

Out of the twenty-eight patients, thirteen were males (47%) and fifteen were females (53%). Four polyps, ten erosions, eight ulcers, two bleeding regions, one stricture, two worms and one angioectasia were observed in twenty-three videos. Five videos were found normal. The average small bowel passage time was 279.6 minutes (SD=110.2). Figure 4.5 depicts the patient demographics of the study.

4.5.2 AI-KODA Score Application Feedback and Efficacy

All the three gastroenterology fellows rated the AI-KODA score application user experience as excellent in the feedback form. All of them favored to utilize the application in their real-time practice with suggested modifications like inclusion of more examples in training module for clarification of ambiguous frames, automatic stool and bubble identification, and automatic deletion of the repetitive frames per video. The fellows took about 2-10 seconds time (SD=2.05) to score each frame and referred to the reference sheet for about 2-20 times (SD= 8.65) during the entire study.

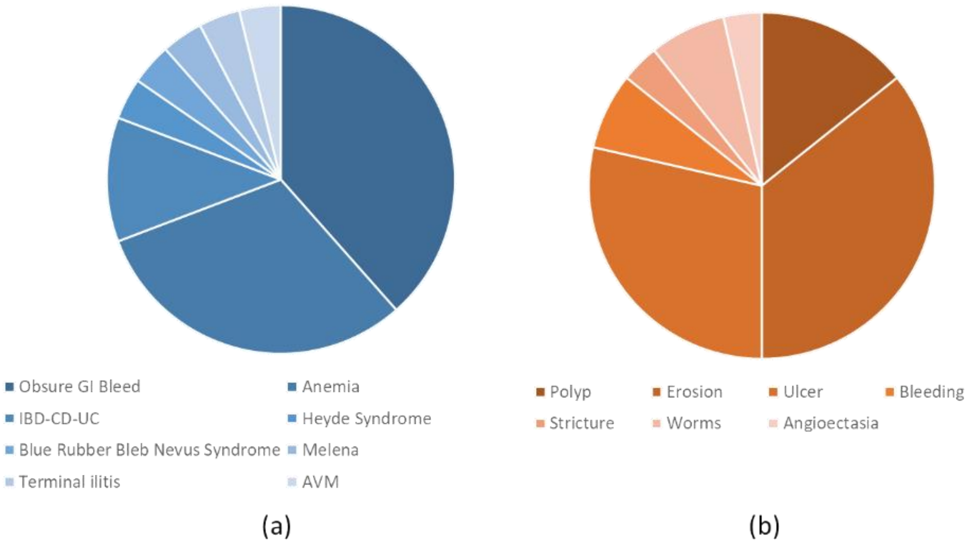


Figure 4.5 Demographics of the twenty-eight capsule videos. (a) Types of diseases diagnosed, (b) Types of anomalies found.

4.5.3 Reliability estimates

Inter-rater and intra-rater estimates for the five-minute interval frames and random-interval frames are summarized in Table 4.1 and 4.2.

Table 4.1 Average measure of inter-rater intra-class correlation coefficients for five-minute interval and random intervals in twenty-eight capsule videos.

Average measure	Five-minutes interval			Random interval		
	Cycle 1	**Cycle 2**	**Ar. Mean**	**Cycle 1**	**Cycle 2**	**Ar. Mean**
	(a) Sum of sub-score 1			(a) Sum of sub-score 1		
Intraclass Correlation	0.96	0.98	0.97	0.99	0.98	0.99
Lower bound	0.92	0.97	0.94	0.99	0.97	0.98
Upper bound	0.98	0.99	0.98	0.99	0.99	0.99
	(b) Sum of sub-score 2			(b) Sum of sub-score 2		
Intraclass Correlation	0.91	0.97	0.94	0.99	0.97	0.98
Lower bound	0.84	0.95	0.89	0.98	0.95	0.97
Upper bound	0.95	0.98	0.97	0.99	0.98	0.99
	(c) Sum of final score			(c) Sum of final score		
Intraclass Correlation	0.84	0.83	0.84	0.91	0.89	0.90
Lower bound	0.70	0.69	0.70	0.84	0.80	0.82
Upper bound	0.92	0.91	0.92	0.95	0.94	0.95

For sequential frames, ICCs for inter-rater variability of sum of sub-score 1 (ICC 0.97, 95 % CI 0.94-0.98), sum of sub-score 2 (ICC 0.94, 95 % CI 0.89-0.97), and final score

(ICC 0.84, 95 % CI 0.70-0.92) were excellent to good among the three fellows. For random frames, ICCs for inter-rater variability of sum of sub-score 1 (ICC 0.99, 95 % CI 0.98-0.99), sum of sub-score 2 (ICC 0.98, 95 % CI 0.97-0.99), and final score (ICC 0.90, 95 % CI 0.82-0.95) were excellent among the three fellows. In sequential frames, ICCs for intra-rater variability of sum of sub-score 1 (ICC 0.92, 95 % CI 0.83-0.96), sum of sub-score 2 (ICC 0.84, 95 % CI 0.66-0.92), and final score (ICC 0.52, 95 % CI 0.29-0.78) were good to moderate among the three fellows. For random frames, ICCs for intra-rater variability of sum of sub-score 1 (ICC 0.98, 95 % CI 0.96-0.99), sum of sub-score 2 (ICC 0.97, 95 % CI 0.95-0.98), and final score (ICC 0.90, 0.78 % CI 0.78-0.95) were excellent among the three fellows.

Table 4.2 Average measure of intra-rater intra-class correlation coefficients for five-minute interval and random intervals in twenty-eight capsule videos.

Average measure	**Five-minutes interval**				**Random interval**			
	Doctor 1	**Doctor 2**	**Doctor 3**	**Ar. Mean**	**Doctor 1**	**Doctor 2**	**Doctor 3**	**Ar. Mean**
	(a) Sum of sub-score 1				(a) Sum of sub-score 1			
Intraclass Correlation	0.97	0.95	0.84	0.92	0.99	0.96	0.99	0.98
Lower bound	0.94	0.89	0.66	0.83	0.98	0.92	0.98	0.96
Upper bound	0.98	0.97	0.92	0.96	0.996	0.98	0.99	0.99
	(b) Sum of sub-score 2				(b) Sum of sub-score 2			
Intraclass Correlation	0.96	0.91	0.65	0.84	0.99	0.94	0.99	0.97
Lower bound	0.91	0.82	0.25	0.66	0.99	0.88	0.98	0.95
Upper bound	0.98	0.96	0.84	0.92	0.99	0.97	0.99	0.98
	(c) Sum of final score				(c) Sum of final score			
Intraclass Correlation	0.80	0.67	0.09	0.52	0.92	0.86	0.91	0.90
Lower bound	0.57	0.30	0	0.29	0.84	0.70	0.81	0.78
Upper bound	0.90	0.85	0.58	0.78	0.96	0.93	0.96	0.95

4.6 Discussion

We developed a simple, easy-to-use mobile-based application to automate the task of scoring the VCE frames as per existing KODA scoring system for an effective cleanliness assessment in VCE and determined its inter-rater and intra-rater reliability.

To the best of the author’s knowledge, this is the first attempt to automate this process. Through this process, we were able to develop a multi-label dataset for prospective AI applications in this field. Also, ours is the third study to show the effectiveness of KODA for cleanliness assessment in research and clinical use. The achieved results for inter-rater and intra-rater reliability were found overall good. They were consistent with the original KODA study for inter-rater ICCs of five-minute intervals.

We agree that the existing KODA score is simple and has a face validity as claimed by the original KODA study. The analysis of visualized mucosa helps a gastroenterologist to analyze the examination quality and the likelihood of missing the lesion according to bowel preparation degree [145]. This analysis is an important indicator of intra-procedural quality on VCE [145]. However, there are some inherent challenges while scoring the frames as per the second question i.e., percentage of obstructed view. We observed that the intra-rater reliability of sum of sub-score 2 (percentage of obstructed view) declined in case of gastroenterology fellow 2 and 3. Similarly, due to the decline in the sub-score 2 ICCs, the overall score ICCs also declined drastically for intra-rater estimates. Slight decline was noticed in inter-rater ICCs for sum of sub-score 2. This is possibly due to lack of examples in the training module and their associated instructions for frames containing shadows and different types of obstructions like fluids, bubbles, solid food and fibers etc. In our opinion, the existing KODA score faces a practical knowledge gap for its successful implementation in real-time clinical settings. There is a need for the inclusion of more frames in the training module which focus on addressing these challenges. Some ambiguous frames have been discussed in Figure 4.6.

Presently, about 3-4 hours are consumed per video analysis for manual anomaly detection in VCE [146], [147]. During this duration, the VCE cleanliness preparation is also assessed and reported by gastroenterologists. So, to estimate the efficacy of the developed application, we assumed that if a reader takes about thirty seconds to score one frame using the developed application, then 2173 frames extracted from 28 capsule videos will be completed in about 22.35 hours. Hence, per video analysis for a small bowel transit time of four hours, in a real-time setting using the developed application will take about 24 minutes. In the study, the gastroenterologists scored one frame in between 2-10 seconds (SD= 2.05) and took about 1.6 minutes - 8 minutes per video

(SD= 4.62). Thus, showing the time-saving efficacy of the developed application in comparison to the manual KODA score which may require use of pen and paper or self-calculations during analysis. This will be further reduced with the help of AI techniques and automatic frame extraction to fully automate the cleanliness assessment of VCE videos in real-time.

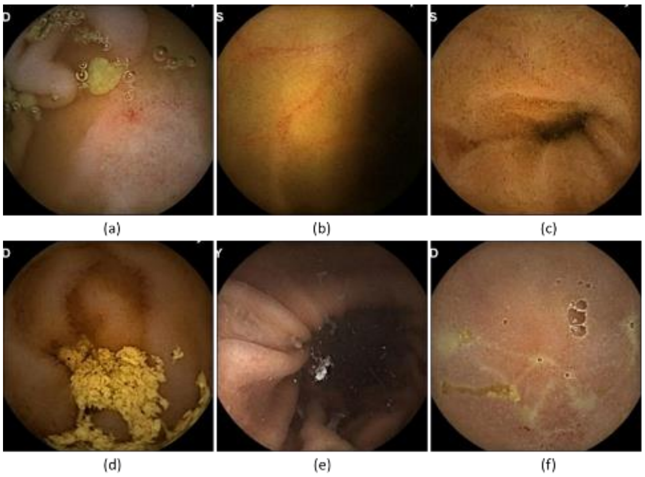


Figure 4.6 (a) – (f) Sample frames with ambiguous scores. In case of (a), the opinion varied between <5% and 5%-25% for percentage of obstructed view. In case of (b), the opinion varied between 5%-25% and 26%-50% for percentage of obstructed view. In (c), the opinion varied between <5% and 5-25% for percentage of obstructed view. Some ambiguities were found for the percentage of mucosa visualized. In (d) the opinion varied between >75% and 25%-49% for the percentage of mucosa visualized. Ambiguous scores of percentages of mucosa visualized and obstructed view were achieved in both (e) and (f). For (f), the opinion varied between 50%-75% and 25%-49% for mucosa visualized and 5%-25% and >50% for obstructed view. Similarly, in (f) the opinion varied between <25% and 50%-75% for mucosa visualized and >50% and <5% for obstructed view.

The previous scores have considered different types of intervals like two-minute, first and last ten-minutes of the small bowel segments, first five-minutes of each segment, random intervals etc., while selecting the VCE frames for scoring purpose [101], [138], [139], [141], [148]. The original KODA score considers frame selection at an interval of five-minutes [140]. In this study, two types of intervals (five-minutes and random) were considered while assessing the frames as per existing KODA to check the effect of intervals with respect to its scoring system. It was observed that the ICCs showed a similar trend as observed in the five-minute interval and were found reliable for both inter-rater and intra-rater estimates. This analysis was also done for the future development of an AI based scoring system which will be potentially free-of intervals effect, assess the entire video and represent a generalized scoring system for cleanliness assessment in VCE. Thus, solving the consideration of intervals in existing scoring systems.

There are two main limitations to the developed application i.e., it is presently available only for Android device users and users must login from the same Android device while scoring the frames. The history information was not saved to maintain the confidentiality of the frames and the users.

4.7 Conclusion and Future Scope

In conclusion, the developed application automates the process of scoring the VCE frames as per the existing KODA score which saves time in cleanliness assessment and is user-friendly for research and clinical use. The achieved inter-rater estimates are encouraging. Further research on the existing KODA scoring system is required to achieve an improved intra-rater reliability. The development of an automatic multi-label cleanliness assessment for VCE with the help of the developed data will be discussed in the next chapter.

CHAPTER 5

DEVELOPMENT OF CLASSIFICATION TECHNIQUES FOR CLEANLINESS ASSESSMENT IN ENDOSCOPY

5.1 Introduction

In continuation to the discussed limitations in the previous chapter, the limitations can be combatted with the development of reliable and efficient computer-aided automatic scoring systems for health experts in gastroenterology department using AI algorithms. Automatic computer-operated assessments have been introduced in this field to automate and reduce the assessment time of cleanliness in VCE [106], [137], [149], [150]. These studies may prove to be more reliable, quick for assessment, and provide independence from the intra-observer and inter-observer variability.

5.2 Related Works

We discuss the related works in this field, briefly, in two perspective namely computer-assisted cleanliness assessments, and multi-label classification in VCE.

Klein *et al.* [101] designed and validated a computer-aided small bowel preparation score based on the pixels in the colour bar of VCE frames. They categorized the frames into ‘adequate’ and ‘inadequate’ and performed VCE frame classification. Pietri *et al.* [102] focused on developing a computer-aided system based on four different statistical features namely grey-level correlation matrix, speeded up robust features, fractal dimension features, and Hough transform to evaluate the abundance of bubbles in VCE frames. Four hundred still frames were categorized as ‘scarce in bubbles’ or ‘abundant in bubbles’ based on the percentage presence of bubbles observed by the physician in the VCE frames. A sensitivity of up-to 94.74% was achieved on the validation set. Ali *et al.* [103] developed a CAC score to inspect the quality of VCE still frames as ‘adequate’ and ‘inadequate’. The authors used channels of colour intensities of the RGB model and extracted it for each frame. A CAC score cut-off of 1.6 validated a sensitivity of up to 91.3% and a specificity of up to 94.7% for 228 still frames. Oumrani *et al.* [104] developed an automatic rapid tool for assessing mucosal visualization quality of still VCE frames using colour intensity ratio, brightness index and grey level correlation matrix features and random forest classifier. Six hundred normal still VCE frames were extracted and evaluated through ten-point assessment grid. The combination of the mentioned features produced a sensitivity up-to 90%.

Noorda *et al.* [74] developed a deep learning algorithm with light weight and reduced trainable parameters to automatically evaluate the degree of cleanliness in VCE on an intuitive scale namely ‘clean’ or ‘dirty’. To locate and quantify the intestinal content, the authors developed patches from the extracted frames. Nam *et al.* [151] developed a deep learning-based software to calculate cleansing score in VCE. They used 700 frames per each cleansing score and implemented an existing deep learning model named Inception residual network version two. Abnormalities such as polyp, ulcer, bleeding etc., found in the frame were not considered in the training and test set. The top $-$ 1 and top $-$ 2 accuracies achieved by the network were 69.4% and 91.2%, respectively. Based on the classification, scores were manually assigned and compared with physician’s decision to check the efficacy of the network. A similar work by the authors used generic convolutional neural network consisting of five convolutional and max pooling layers with one full connected layer for 4,00,000 still frames categorized into score 1 $-$ 5 depending on mucosal visibility [152]. Their network achieved an accuracy up-to 93% on 120 test set frames and misclassification rate up-to 24.7% on 51,380 separate set of VCE frames. A neural network based automatic cleanliness scoring was proposed using six hundred normal still VCE frames [153]. The frames were categorized as ‘adequate’ and ‘inadequate’ based on ten-point scale. The authors reported an accuracy up-to 89.7%.

Now we discuss multi-label classification in VCE which has been scarcely explored. Vasilakakis *et al.* [107] investigated the semantics of a VCE video content and developed a multi-label classification method for five categories namely abnormal, debris, bubble and lumen hole. The authors utilized bag-of-words approach on CIE-Lab converted VCE RGB frames and a convolutional neural network architecture enabling multi-scale feature extraction to categorize the multi-labels. Ratsnake software was utilized to add the multi-labels. The authors did not consider dependencies between la- bels and associated cleansing score. The developed method achieved an area under curve (AUC) score up-to 0.94, 0.91 and 0.85 for debris, bubble and lumen hole classes. Park and Lee [108] proposed a class-labelling method that may be used to design a learning model by constructing a knowledge model focused on main lesions defined in standard terminologies for VCE such as minimal standard terminology and VCE structured terminology. The knowledge model considered the anatomy of the GI tract and findings in VCE. Three major class labels namely normal, abnormal and discriminative class were given. The normal class labels were further distinguished based on the bubbles, wrinkles, and location of the capsule. The discriminative classes contained frames due to low power, transmission or reception problems, and a large amount of foam. The special cases were analyzed by developing clusters of colour similarity through k-means algorithm. The supra- and sub-classes were made using the concept of ontology. The authors conducted a classification task to distinguish the different organs using a generic convolutional neural network and achieved an accuracy up-to 33.5%. Mohammed *et al.* [109], [154] developed a pathology-sensitive abnormality detection through deep learning algorithms for colon diseases in VCE data. They developed VCE dataset (presently private) containing 455 short video segments with 28,304 frames and 14 classes of colorectal diseases. The classes consisted of abnormalities such as erosions, debris, diverticulosis, erythema, granularity, haemorrhage, inflammation, edema, angioectasia, polyp, pseudo polyp, tumour and ulceration. The authors performed video and frame-level prediction and achieved an average precision, recall, F1-score and specificity up-to 61.6%, 54.6%, 55.1%, and 95.1% respectively.

Table 5.1 discusses the previous approaches in this field. In particular, the studies have compared the red over green pixel ratio (R/G), R/G ratio, brightness, abundance of bubbles, and checking adequacy and in- adequacy using neural networks; they have also looked at automatic score classification of visible mucosa, dirty and clean patches, and colour bar differentiation.

Table 5.1 Previous approaches in this field.

Ref. and Year	**Type of classification**	**Description**
Klein *et al.* [101] and 2016		Designed and validated a computer-aided small bowel preparation score based on the pixels in

	Binary (adequate v/s in-adequate)	the colour bar of VCE frames.
Pietri *et al.* [102]and 2018	Binary (bubble v/s no bubble)	Focus on evaluating abundance of bubbles in VCE frames. Used 400 frames. Sensitivity achieved up to 94.74%
Ali *et al.* [102] and 2018	Binary (adequate v/s in-adequate)	Computer-aided assessment of cleansing score for checking sufficiency of VCE frames. Used 228 frames. Sensitivity achieved up to 91.3%.
Oumrani *et al.* [104] and 2019	Binary (adequate v/s in-adequate)	Automatic rapid tool for assessing mucosal visualization quality in VCE frames. Used 600 frames. Sensitivity achieved up-to 90%.
Noorda *et al.* [74] and 2020	Binary (clean v/s dirty)	Automatic evaluation of degree of cleanliness in VCE. Used 700 frames as patches. Sensitivity achieved up-to 91.2%
Leenhardt *et al.* [153]and 2021	Binary (adequate v/s in-adequate)	Neural network for automatic cleanliness scoring in VCE. Used 600 frames. Accuracy achieved up-to 89.7%.
Nam *et al.* [151]and 2021	Multi-class (1 v/s 2 v/s 3 v/s 4 v/s 5)	Software to calculate cleansing score in VCE. Used 4L frames in training. Removed abnormal frames. Accuracy up-to 93% for 120 frames. 1-5 scales are as per adequacy.
Mohammed *et al.* [109] and 2020	Multi-class	Developed a pathology sensitive abnormality detection in VCE for colon diseases. Used 28, 304 frames and 14 classes. Precision achieved up-to 54.6%.
Vasilakakis *et al.*[107] and 2020	Multi-label	Investigation in semantics of VCE video content. Categories are abnormal, debris, bubble and lumen hole. Ratsnake software was utilized to add the multi-labels. AUC achieved between 85 − 94%.
Park and Lee [108]and 2020	Hierarchical	Class-labelling method for lesion detection as minimal standard terminology in VCE. Supra classes included normal, abnormal, and discriminative. Sub-class included bubbles, wrinkles, and location of the capsule. Accuracy achieved up-to 33.5%.

It can be noticed that most studies have been designed for CE frames up to 700 with binary class labels such as ‘clean’ and ‘dirty’, ‘adequate’ and ‘inadequate’. No computer-operated assessment study to automate an existing medical score for cleanliness assessment in VCE has not been done. The same conclusion was also reported in [137]. The primary reason behind this is the absence of high quality, multi-labelled, and medically validated AI dataset. Lack of data creates a serious limitation on the performance and re- producibility of the existing computer-operated assessments in this field. Automatic multi-label classification is an emerging and presently less explored area in this field. It has the potential to address automatic scoring system for cleanliness assessment in VCE.

CE cleanliness is the only dataset which is publicly available for assessing the grade of cleanliness. It was released in 2020 and acquired at Hospital Universitari i Politècnic La Fe from Valencia, using Pillcam SB 3 system. 563 individual frames of 576 × 576 pixels were extracted from 35 different CE videos during patient procedures and considered as training set. 854 additional frames of 576 × 576 pixels were extracted from 30 additional CE videos of different patients for development of validation set. The dataset has been built to locate and quantify the intestinal content in a CE procedure where-in the extracted frames have been cut into patches of 64 × 64 pixels, with a step size of 32 pixels which are class labelled as dirty or clean. No other medical information related to the findings of the CE and their class labels has been mentioned in the dataset.

5.3 Problem formulation

The primary objective of the present study was to develop an automated approach to automatically assess the cleanliness of VCE as per the latest scoring system using AI models in real-time. The secondary objective of the present study was:

- To develop a simple and user-friendly application for gastroenterologists to score the VCE frames as per the latest scoring system i.e., KODA. The application is called as AI-KODA score. It is fully auto- mated and works in real-time on an Android phone. It also helps in development of a multi-labelled image dataset.
- To generate a high-quality, multi-labelled image dataset which is medically validated and can be utilized for development of AI models for computer- operated assessments in this field.
- To conduct a comprehensive evaluation, interpretation, benchmarking of the generated dataset using famous AI algorithms.

We have achieved the first objective in this previous chapter. Now the second and third objective will be discussed in this chapter. For this, the problem formulation for automatic classification of the developed multi-labelled dataset will be discussed. The dataset consists of the same 2173 VCE frames.

Let $X \in R$, $i \times j \times k$ represents the tensor of an RGB-coloured VCE frame wherein i, j, and k are the frame width, height, and colour channels, respectively. For each frame, there is an output tensor Y of labels q, $q > 1$. The output tensor consists of two label sets L1 and L2. Both the label sets are distinct. The labels in each of the label set are mutually exclusive. Their probability of occurrence is 1/4. For automatic classification purpose, X_m is reshaped to a one-dimensional (1-D) vector of n features. Similarly, Y_m is reshaped to 1-D array with indices of the max element of the array in a particular axis. m represents the no. of frames in this study (2173). Then X_m and Y_m are randomly split into training, validation, and testing ratio of 70%, 20%, and 10% respectively which are used for fitting and predicting values from each classifier.

5.4 Methodology

After the scoring of the frames, two times, the frames with common sub-scores were extracted. The frames with ambiguous sub-scores were sent back to the gastroenterology fellows for a third-time review. The final decision was made with the help of a senior gastroenterologist. Finally, a multi-label image dataset with eight labels (VM and OV sub-scores) was developed. It was subjected to exploratory data analysis (EDA). Then the dataset was transformed into three datasets namely ‘only VM scores’, ‘only OV scores’, and ‘both VM and OV scores. Figure. 5.1 depicts the label distribution in the dataset. Clearly, it is an un-balanced dataset. Highest no. of labels in VM and OV were VMscore3 and OVscore2 respectively. Lowest no. of labels in VM and OV were VMscore1 and OVscore3 respectively. The correlation between the labels has been shown in Figure 5.2. VM sub-scores were negatively correlated with each other. A similar trend can be observed in OV sub-scores. OVscore1 and VMscore0 were highly positively correlating with each other.

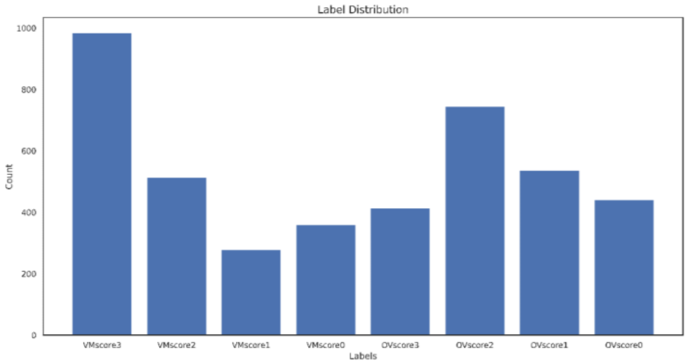
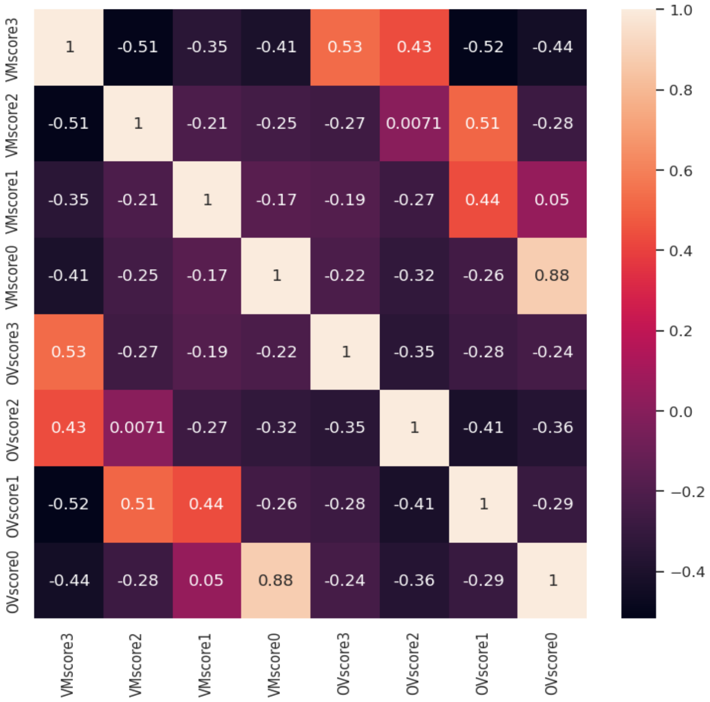


Figure 5.1 Label distribution of the developed dataset.

For machine learning experiments, each of the dataset was then randomly split into train: validate: test ratio of 70: 20: 10. After the random split, in each of the dataset, 1521 frames represented the train data, 434 represented the validation data, and 218 frames represented the test data. A comprehensive evaluation was done for the three datasets using ten famous machine learning algorithms. The machine learning algorithms used were RF classifier, ridge classifier, bagging classifier, multi-layer perceptron (MLP) classifier, KNeighbours classifier, decision tree (DT), support vector classifier (SVC), Gaussian naive Bneuayes (NB), logistic regression (LR), and Adaboost classifier.

For transfer learning experiments, each of the dataset was then randomly split into train: validate: test ratio of 60:20:20. They were split in a manner that there is no repetition of frames nor data leakage in any of the training, validation, and testing datasets. A comprehensive benchmarking and evaluation were done using eight famous transfer learning algorithms namely VGG 19, ResNet 50 V2, ResNet 152V2, Inception v3, Inception residual network, xception, mobile network (MobileNet) V2, and dense network (DenseNet) 169 base networks for each of the dataset. We discuss briefly about the machine learning and transfer learning algorithms.



	VMscore3	VMscore2	VMscore1	VMscore0	OVscore3	OVscore2	OVscore1	OVscore0
VMscore3	1	-0.51	-0.35	-0.41	0.53	0.43	-0.52	-0.44
VMscore2	-0.51	1	-0.21	-0.25	-0.27	0.0071	0.51	-0.28
VMscore1	-0.35	-0.21	1	-0.17	-0.19	-0.27	0.44	0.05
VMscore0	-0.41	-0.25	-0.17	1	-0.22	-0.32	-0.26	0.88
OVscore3	0.53	-0.27	-0.19	-0.22	1	-0.35	-0.28	-0.24
OVscore2	0.43	0.0071	-0.27	-0.32	-0.35	1	-0.41	-0.36
OVscore1	-0.52	0.51	0.44	-0.26	-0.28	-0.41	1	-0.29
OVscore0	-0.44	-0.28	0.05	0.88	-0.24	-0.36	-0.29	1

Figure 5.2 Correlation matrix of the labels.

RF is a supervised machine learning algorithm used for both classification as well as regression tasks [155]. The concept of RF is based on ensemble learning, a process of combining multiple classifiers to improve the accuracy of the model [155], [156]. RF employs several decision tree classifiers on the dataset and averages the result across all the trees to output a result for regression, meanwhile, the most commonly predicted label is chosen for classification [155], [156].

A DT is a machine learning algorithm with a treelike structure, depicting all possible outcomes for a particular choice. It is employed for both classification and regression and is a part of the supervised branch of machine learning algorithms, i.e. the model is trained and tested on a set of data containing the desired classification result. The base of the DT is a root node containing several branches, following which are decision nodes representing the decisions that are to be made, leaf nodes depict the outcome of those decisions. The DT keeps growing by recursively splitting based on the attributes of the training data until a stopping criterion is met [156], [157].

Ridge classifier is a machine learning algorithm designed for binary and multi-class classification tasks [156]. By combining ideas from conventional classification techniques and Ridge Regression, it offers a distinct method for classifying data points.

Bootstrap aggregation (bagging) is an ensembling method that attempts to resolve overfitting for classification or regression problems [155]. Bagging aims to improve the accuracy and performance of machine learning algorithms. Each classifier's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set. The predictions for each subset are then aggregated through majority vote for classification or averaging for regression, increasing prediction accuracy.

In machine learning, linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables) [155], [156]. It falls under the category of regression algorithms, which are used when the target variable is a real or continuous value. Researchers have used it for classification problems as well.

LR is a statistical method commonly used in machine learning for binary classification problems. Despite its name, it is a classification algorithm rather than a regression algorithm. LR predicts the probability that an instance belongs to a particular category. For example, email spam or not. The output is a value between 0 and 1, which represents the probability of the instance belonging to the positive class. It is a supervised machine learning algorithm. It can be extended to handle multi-class classification as well [155].

The Gaussian Naive Algorithm is a form of the Naive Bayes algorithm used in classification problems. The Naive Bayes algorithms are a collection of algorithms used in machine learning. The is a probabilistic algorithm that all the features in a class are independent of each other. In Gaussian Naive Bayes the features of the data are continuous and have a gaussian distribution. This algorithm works on continuous and normally distributed features [155].

The SVC is an integral element within the expansive domain of machine learning, introduced in the early 1960s [155]. This historical foundation laid the groundwork for SVC, solidifying its position as a cornerstone within the discipline. Motivated by the ambition to devise robust classification algorithms capable of navigating intricate decision boundaries with unparalleled accuracy, SVC naturally evolved as an extension of Vapnik and Chervonenkis' pioneering work on SVM [156]. The impetus behind SVC is firmly rooted in the pursuit of a classification model possessing both formidable power and versatility, ensuring high accuracy across diverse datasets.

The nomenclature of the Support Vector Classifier pays homage to the central concept of "support vectors" inherent in SVM. In SVM-based classification, these support vectors exert significant influence over the determination of the decision boundary's positioning and orientation [157]. This conceptual foundation serves as the bedrock upon which SVC establishes its classification prowess, emphasizing its continuity with the principles set forth by SVM.

The mathematical underpinnings of SVC closely echo the principles articulated by SVM [155]. At its core, the essence of SVC lies in the determination of an optimal hyperplane that maximally segregates different classes within the feature space. Articulated as an optimization problem, SVC is designed to maximize the margin between classes while simultaneously minimizing classification errors, endowing it with robust discriminative capabilities.

SVC's efficacy transcends disciplinary boundaries, rendering it a versatile workhorse in a myriad of classification tasks [155]. Particularly adept in scenarios characterized by non-linear and complex decision boundaries, SVC finds applications across a broad spectrum, spanning domains such as image classification, text categorization, bioinformatics, and beyond. The adaptability of SVC positions it as an indispensable tool in various fields where precise and accurate classification holds paramount importance.

Gradient Boosting, a transformative concept in machine learning, represents an evolution rather than a singular model, with its development unfolding over the course of time [155]. The foundational work on gradient boosting was elucidated by Jerome

H. Friedman in the late 1990s, marking the Inception of a paradigm that has since become instrumental in enhancing predictive accuracy [156], [157]. His seminal work laid the groundwork for a technique that redefined the landscape of ensemble learning.

At the core of Gradient Boosting lies an inspiration to augment the accuracy of predictive models by synergistically combining the strengths of multiple weak learners, typically in the form of decision trees. The motivation is to harness the collective power of these learners in an additive manner, mitigating their individual weaknesses.

The nomenclature "Gradient Boosting" is a nod to the optimization technique of gradient descent employed in the process. The term "boosting" encapsulates the iterative approach of enhancing model performance by sequentially adding weak learners to rectify the errors of their predecessors. This iterative refinement process forms the crux of the technique.

The mathematical foundation of Gradient Boosting revolves around the minimization of a loss function through the application of gradient descent [155]. In each iteration, a weak learner is introduced into the ensemble to correct errors made by the existing model. The weights of misclassified instances are strategically adjusted to guide the ensemble towards the optimal solution, creating a powerful and adaptive model.

The architecture of Gradient Boosting manifests in its iterative, additive nature. At its core, it assembles an ensemble of weak learners, often decision trees, sequentially. In each iteration, a new tree is created to correct the errors of the ensemble up to that point. Combining these weak learners leads to a robust and accurate predictive model.

Gradient Boosting, known for its versatility, finds applications in various fields, including regression and classification problems. Implementations like XGBoost, LightGBM, and AdaBoost have emerged as stalwarts, showcasing exemplary performance in many domains like finance, healthcare, and natural language processing [155], [156]. This adaptability positions Gradient Boosting as a go-to technique in diverse and complex problem-solving scenarios.

The VGG model often called as VGGNet is a convolutional neural network model proposed by A. Zisserman and K. Simonyan of the University of Oxford in 2014 that supports 16 layers [117]. The VGG architecture serves as a foundation for cutting edge object recognition models. It was created as a deep neural network, outperforms baselines on a variety of tasks and datasets into ImageNet [117], [119]. It is usually pre-trained on the ImageNet dataset, containing millions of images across thousands of categories. This pre-training helps the model learn a rich set of hierarchical features. After pre-training, the model can be fine-tuned on specific datasets or used as a feature extractor for various image-related tasks through transfer learning.

ResNet50V2, an evolution born from the pioneering ResNet50, represents a significant stride in the realm of deep learning architectures [117]. The original ResNetV2 paper was presented in 2016, signalling a deliberate effort by researchers from Microsoft Research, including Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, to enhance the ResNet lineage [117]. The inspiration behind ResNet50V2 stems from a concerted effort to refine the original ResNet architecture. Recognizing the need for improvements in training efficiency and model generalization, ResNet50V2 emerges as a testament to the continuous pursuit of excellence in deep learning.

The nomenclature "ResNet50V2" carries the implication of version improvement over the original ResNet, denoted by the "V2" suffix. This signifies the incorporation of advancements aimed at addressing limitations and elevating overall model performance. While maintaining the fundamental structure of ResNet50, ResNet50V2 introduces notable tweaks in its architecture. A pivotal change involves the adoption of "bottleneck residual blocks" accompanied by additional batch normalization before each weight layer. These modifications are strategically designed to enhance model generalization and facilitate more seamless optimization during training. The mathematical principles governing ResNet50V2 closely align with those of its predecessor, ResNet50 [117]. The incorporation of additional batch normalization steps remains a key strategy, aimed at facilitating the training of deep networks by ensuring a more efficient gradient flow. These principles underscore the model's commitment to addressing challenges in training deep neural networks.

ResNet50V2 finds its stride in a multitude of computer vision applications, mirroring the domains where ResNet50 has excelled [117]. The enhancements in training efficiency and generalization position ResNet50V2 as a preferred choice in scenarios where heightened performance is crucial. The model's versatility extends its utility beyond research labs, making it an asset in industries relying on computer vision for tasks such as image recognition, object detection, and image segmentation.

The Inception residual network was created by Christian Szegedy, Sergey Ioffe and Vincent Vanhoucke in the year 2016 [117]. The model is basically an extension of the Inception network and is made using a culmination of the Inception infrastructure and Residual Connections which explains its name Inception Residual Network or Inception Resnet. Inception basically applies various transformation operations to the input data and gives a concatenated output. They are great at creating deep models that are still computationally efficient.

A problem faced in training deep neural networks is the vanishing gradient problem [117], [119], [120]. As back propagation occurs through the many layers of a deep neural network, gradients (which are loss function derivatives for network parameters) start becoming increasingly small. This is not ideal as it subdues the training process. To solve this issue residual connection, or skip connections, were introduced to the Inception block giving this model high feature-extraction power with a built-in fail-safe to gradient vanishing for deep models [117]. Now we will discuss rest of the followed methodology to evaluate and run the discussed machine learning and transfer learning algorithms.

The evaluation metrics was achieved using predicted scores by machine learning or transfer learning algorithms v/s the true scores given by the readers. The metrics used for the three multi-label machine learning-based classification tasks included overall accuracy, balanced accuracy, weighted average of the achieved precision, recall, F1-score, and Jaccard score. The metrics were reported for both the validation and test data. Overall accuracy, binary accuracy, overall categorical accuracy, loss, precision, recall, and top-k categorical accuracy were used as the evaluation metrics for the three multi-label transfer learning-based classification tasks for both training and validation data. Weighted average of the achieved precision, recall, and F1-score were reported for the testing data. Some of the formula for the utilized evaluation metrics have been mentioned below:

Classification accuracy is the percentage of properly defined data from the complete set indicated by the TP and TN condition or the percentage ratio of correctly specified data to the full dataset, as shown in Eq. (5.1).

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \qquad (5.1)$$

Here, TP is number of true positives, TN is number of true negatives, FP is number of false positives, and FN is number of false negatives.

Precision is the uniformity of the measuring findings as shown in Eq. (5.2).

$$Precision = \frac{T_P}{T_P + F_P} \qquad (5.2)$$

Recall is the proportion of similar occurrences recovered as shown in Eq. (5.3).

$$Recall = \frac{T_P}{T_P + F_N} \qquad (5.3)$$

F-score is the weighted harmonic mean of precision and recall as shown in Eq. (5.4).

$$F - score = \frac{Precision \times Recall}{Precision + Recall} \qquad (5.4)$$

The machine learning-based experiments were run on Google Co laboratory with Python 3 Google Compute Engine backend (GPU). For all the classification tasks, a hyper image dimension, channels, and mode of 64×64, 3, and RGB respectively were set. All the images were normalized between pixel values of 0 and 1. A random seed was initialized to obtain reproducible tensors in the appropriate code. No other hyper-parameter tuning, nor augmentation techniques were utilized to report the results of vanilla machine learning algorithms.

The benchmarking setup of transfer learning experiments were developed on Python with TensorFlow in the back end on an Intel(R) Xeon(R) Silver 4214 CPU @ 2.20

GHz with 2 processors, 128 GB RAM, and dedicated 24 GB NVIDIA RTX A5000 workstation. For all the classification tasks, a hyper image dimension, batch size, epochs, channels, and mode of 320 × 320, 32, 250, 3, and RGB respectively were set. All the images were normalized between pixel values of 0 and 1. No other augmentation techniques were utilized. The Adam algorithm was used as the optimizer with a learning rate of $1e-3$, a beta 1 of 0.9, a beta 2 of 0.999, epsilon of $1e-7$, and loss as binary cross-entropy. A random seed was initialized to obtain reproducible tensors in the appropriate code with weights taken from ImageNet for all the transfer learning algorithms. The last layer was activated using a sigmoid function to predict the probability of each label between 0 and 1. Trainable, non-trainable, and total parameters of the transfer learning algorithms along with their size have been mentioned in Table 5.2.

Table 5.2 Details of the transfer learning algorithms used in the three multi-label classification tasks.

S. No.	**Transfer learning algorithm**	**Trainable parameters**	**Non-trainable parameters**	**Total parameters**	**Size**
		(a) Only VM labels and only OV labels			
1.	VGG 19	204,804	20,024,384	20,229,188	78.8 MB
2.	ResNet50V2	819,204	23,564,800	24,384,004	99.7 MB
3.	Inception v3	524,292	21,802,784	22,327,076	90 MB
4.	InceptionResNetV2	393,220	54,336,736	54,729,956	213 MB
5.	Xception	819,204	20,861,480	21,680,684	89.3 MB
6.	MobileNetV2	512,004	2,257,984	2,769,988	14.9 MB
7.	DenseNet 169	665,604	12,642,880	13,308,484	57.4 MB
8.	ResNet152V2	819,204	58,331,648	59,150,852	233 MB
		(b) Both VM and OV labels			
1.	VGG 19	409,608	20,024,384	20,433,992	81.1 MB
2.	ResNet50V2	1,638,408	23,564,800	25,203,208	109 MB
3.	Inception v3	1,048,584	21,802,784	22,851,368	96 MB
4.	InceptionResNetV2	786,440	54,336,736	55,123,176	218 MB
5.	Xception	1,638,408	20,861,480	22,499,888	98.7 MB
6.	MobileNetV2	1,024,008	64,097,687	3,281,992	20 MB
7.	DenseNet 169	1,331,208	12,642,880	13,974,088	65 MB
8.	ResNet152V2	1,638,408	58,331,648	59,970,056	242 MB

5.5 Results and Discussion

Achieved evaluation metrics for validation data and test data have been summarized in Table 5.3 and 5.4. In case of only VM labels based multi-label classification, the evaluation metrics ranged between 45.31 − 61.75% for validation data and 40.69 − 57.07% for test data. In case of only OV labels, the evaluation metrics ranged between 42.33 − 59.90% for validation data and 37.33 − 57.88% for test data. In case of both VM and OV labels, the evaluation metrics ranged between 44.08 − 61.05% for validation data, and 45.38 − 62.38% for test data. Overall, for validation and test data, RF outperformed in comparison to the ten machine learning algorithms. The highest overall accuracy, balanced accuracy, weighted average of the achieved F1-score, recall, precision, and Jaccard score was achieved up to 62.38%, 57.26%, 59.93%, 62.38%, 62.58%, and 45.38% respectively.

Table 5.3 Achieved evaluation metrics on validation data.

Type of labels	**Evaluation metrics**	**RF**	**Ridge**	**Bagging**	**MLP**	**KNeighbours**	**DT**	**SVC**	**Gaussian NB**	**LR**	**Adaboost**
Only VM	Overall accuracy	**61.75**	52.99	55.99	57.83	58.29	50.92	53.68	57.60	57.37	58.52
	Balanced accuracy	46.61	41.28	39.78	**47.15**	41.05	40.47	32.53	46.73	45.69	45.87
	F1-score (weighted average)	**58.40**	51.37	52.15	55.92	52.33	51.17	42.26	56.49	55.44	57.84
	Recall (weighted average)	**61.75**	52.99	55.99	57.83	58.29	50.92	53.68	57.60	57.37	58.52
	Precision (weighted average)	57.63	51.20	50.53	55.59	54.49	51.80	43.57	55.89	54.35	**57.86**
	Jaccard score	**45.31**	37.70	39.73	42.11	38.71	37.12	31.07	43.01	41.80	44.32
Only OV	Overall accuracy	**59.90**	40.55	51.38	45.62	43.54	44.70	45.39	52.30	43.54	55.29
	Balanced accuracy	**57.26**	39.99	50.02	38.74	46.27	44.04	37.72	54.42	44.04	54.17
	F1-score (weighted average)	**59.93**	40.49	51.38	37.49	42.83	44.83	36.44	51.50	43.58	55.46
	Recall (weighted average)	**59.90**	40.55	51.38	45.62	43.54	44.70	45.39	52.30	43.54	55.29
	Precision (weighted average)	**62.58**	41.68	53.40	54.75	51.24	45.13	45.08	52.11	43.65	56.83
	Jaccard score	**42.33**	25.43	34.85	24.90	27.40	29.01	24.22	35.16	28.14	38.55
Both VM and OV	Overall accuracy	**61.05**	52.99	58.75	55.06	58.52	48.38	54.60	53.68	56.22	48.15
	Balanced accuracy	**50.82**	43.55	46.86	44.56	35.95	40.44	37.17	46.92	46.29	36.00
	F1-score (weighted average)	**57.71**	51.09	55.36	52.12	54.25	48.97	45.11	53.70	53.55	38.27
	Recall (weighted average)	**61.05**	52.99	58.75	55.06	58.52	48.38	54.60	53.68	56.22	48.15
	Precision (weighted average)	**57.68**	50.13	54.54	51.33	56.94	49.80	46.79	54.22	52.32	37.41
	Jaccard score	**44.08**	36.48	41.33	38.13	39.91	34.55	33.01	39.18	39.63	27.92

Table 5.4 Achieved evaluation metrics on test data.

Type of labels	Evaluation metrics	**RF**	**Ridge**	**Bagging**	**MLP**	**KNeighbours**	**DT**	**SVC**	**Gaussian NB**	**LR**	**Adaboost**
Only VM	Overall accuracy	**56.88**	49.54	47.70	51.83	56.42	44.03	45.87	55.50	51.37	55.50
	Balanced accuracy	45.89	40.93	36.09	44.54	42.77	35.00	31.24	**51.31**	44.15	47.54
	F1-score (weighted average)	52.79	46.26	43.15	49.17	52.53	42.52	33.54	54.66	48.07	**55.40**
	Recall (weighted average)	**56.88**	49.54	47.70	51.83	56.42	44.03	45.87	55.50	51.37	55.50
	Precision (weighted average)	51.43	46.65	42.42	48.75	**57.07**	41.97	40.72	54.30	47.06	56.26
	Jaccard score	38.82	32.10	30.44	35.76	38.05	29.37	23.01	39.35	34.59	**40.69**
Only OV	Overall accuracy	51.83	42.66	48.62	42.20	46.78	46.33	43.11	47.70	46.33	**54.12**
	Balanced accuracy	49.87	42.72	47.16	37.02	49.91	46.80	36.64	49.09	47.92	**53.31**
	F1-score (weighted average)	50.78	42.79	48.45	31.86	46.64	46.38	33.41	46.78	46.06	**54.30**
	Recall (weighted average)	51.83	42.66	48.62	42.20	46.78	46.33	43.11	47.70	46.33	**54.12**
	Precision (weighted average)	55.77	43.37	50.31	38.34	**57.88**	47.34	41.58	46.91	45.99	57.06
	Jaccard score	34.59	27.40	32.13	21.38	30.46	30.29	22.04	31.00	30.35	**37.33**
Both VM and OV	Overall accuracy	61.00	56.88	58.71	58.71	**62.38**	47.24	57.79	58.25	57.79	55.96
	Balanced accuracy	46.51	40.71	44.85	41.27	43.45	37.91	33.10	**49.18**	44.45	35.86
	F1-score (weighted average)	56.97	54.57	56.60	54.77	56.83	49.25	46.56	**59.14**	55.12	46.24
	Recall (weighted average)	61.00	56.88	58.71	58.71	**62.38**	47.24	57.79	58.25	57.79	55.96
	Precision (weighted average)	56.58	53.77	58.43	52.71	57.60	53.10	43.40	**60.39**	55.81	46.68
	Jaccard score	**45.38**	41.88	43.40	42.28	44.01	35.90	35.73	44.79	42.57	35.88

Achieved evaluation metrics for training data, validation data, and test data have been summarized in Table 5.5 and 5.8 for only VM labels. The highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss for an average of 250 epochs during the training stage was achieved up to 0.99, 0.99, 0.98, 1, 0.99, 0.99, and 0.082 respectively. Likewise, 0.99, 0.99, 0.99, 1, 0.99, 0.99, and 0.004 was achieved as the highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss respectively for the last epoch. The highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss for an average of 250 epochs during the validation stage was achieved up to 0.62, 0.82, 0.62, 1, 0.69, 0.57, and 1.37 respectively. Likewise, 0.64, 0.83, 0.62, 1, 0.74, 0.60, and 1.84

was achieved as the highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss respectively for the last epoch. The highest macro average of the achieved precision, recall, and F1- score during the testing stage was achieved up to 0.72, 0.52, and 0.49 respectively. The highest weighted aver- age of the achieved precision, recall, and F1-score during the testing stage was achieved up to 0.71, 0.65, and 0.60 respectively. For the training and validation data, DenseNet169, ResNet50V2, ResNet152V2, and InceptionV3 performed similar. The main difference was observed in the achieved loss value. For the test data, Inception resnet performed the best in comparison to the other algorithms. It was followed by ResNet50V2, ResNet152V2 and then exception algorithm.

Table 5.5 Achieved evaluation metrics on training and validation data for only VM labels.

Evaluation metrics	**VGG 19**	**ResNet 152V2**	**ResNet 50V2**	**MobileNet V2**	**InceptionV3**	**Inception resnet**	**Exception**	**DenseNet169**
For training data								
Overall accuracy (average of 250 epochs)	0.95	0.97	0.97	0.98	0.97	0.98	0.97	0.97
Overall accuracy (last epoch)	0.99	0.97	0.98	0.99	0.98	0.99	0.98	0.98
Binary accuracy (average of 250 epochs)	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Binary accuracy (last epoch)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Overall categorical accuracy (average of 250 epochs)	0.95	0.97	0.97	0.98	0.97	0.98	0.97	0.97
Overall categorical accuracy (last epoch)	0.99	0.97	0.98	0.99	0.98	0.99	0.98	0.98
Top-k categorical accuracy (average of 250 epochs)	1	1	1	1	1	1	1	1
Top-k categorical accuracy (last epoch)	1	1	1	1	1	1	1	1
Precision (average of 250 epochs)	0.95	0.98	0.99	0.98	0.98	0.98	0.98	0.98
Precision (last epoch)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Recall (average of 250 epochs)	0.94	0.98	0.99	0.98	0.98	0.98	0.98	0.98
Recall (last epoch)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Loss (average of 250 epochs)	0.095	0.25	0.096	0.12	0.086	0.083	0.069	0.082
Loss (last epoch)	0.004	0.24	0.065	0.10	0.051	0.050	0.041	0.038
For validation data								
Overall accuracy (average of 250 epochs)	0.55	0.61	0.61	0.56	0.61	0.59	0.60	0.62

Overall accuracy (last epoch)	0.58	0.64	0.60	0.55	0.60	0.58	0.59	0.59
Binary accuracy (average of 250 epochs)	0.75	0.82	0.81	0.79	0.82	0.80	0.80	0.82
Binary accuracy (last epoch)	0.79	0.82	0.81	0.77	0.83	0.80	0.81	0.82
Overall categorical accuracy (average of 250 epochs)	0.55	0.61	0.61	0.56	0.61	0.59	0.60	0.62
Overall categorical accuracy (last epoch)	0.58	0.64	0.60	0.55	0.60	0.58	0.59	0.59
Top-k categorical accuracy (average of 250 epochs)	1	1	1	1	1	1	1	1
Top-k categorical accuracy (last epoch)	1	1	1	1	1	1	1	1
Precision (average of 250 epochs)	0.52	0.69	0.65	0.58	0.69	0.63	0.61	0.68
Precision (last epoch)	0.58	0.65	0.64	0.54	0.74	0.61	0.66	0.64
Recall (average of 250 epochs)	0.53	0.51	0.53	0.54	0.51	0.53	0.55	0.57
Recall (last epoch)	0.52	0.57	0.54	0.56	0.47	0.56	0.46	0.60
Loss (average of 250 epochs)	1.39	6.38	9.12	6.09	3.83	3.51	4.24	3.72
Loss (last epoch)	1.84	7.88	10.31	7.31	5.87	4.16	6.25	5.13

Achieved evaluation metrics for training data, validation data, and test data have been summarized in Table 5.6 and 5.8 for only OV labels. The highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss for an average of 250 epochs during the training stage was achieved up to 0.98, 0.99, 0.98, 1, 0.99, 0.99, and 0.08 respectively. Likewise, 0.99, 1, 0.99, 1, 1, 1, and 0.00 was achieved as the highest overall accuracy, binary accuracy, over- all categorical accuracy, top-k categorical accuracy, precision, recall, and loss respectively for the last epoch. The highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss for an average of 250 epochs during the validation stage was achieved up to 0.56, 0.82, 0.61, 1, 0.69, 0.54, and 1.53 respectively. Likewise, 0.60, 0.83, 0.60, 1, 0.74, 0.63, and 2.25 was achieved as the highest overall accuracy, binary accuracy, over- all categorical accuracy, top-k categorical accuracy, precision, recall, and loss respectively for the last epoch. The highest macro average of the achieved precision, recall, and F1-score during the testing stage was achieved up to 0.56, 0.55, and 0.54 respectively. The highest weighted average of the achieved precision, recall, and F1-score during the testing stage was achieved up to 0.55, 0.60, and 0.54 respectively. For the training and validation data, InceptionV3 performed the best among the eight algorithms; followed by MobileNetV2 and Inception resnet. For the test data, ResNet50V2, Mo- bileNetV2, and DenseNet169 performed similar.

Achieved evaluation metrics for training data, validation data, and test data have been summarized in Table 5.7 and 5.8 for both VM and OV labels. The highest overall

Table 5. 6 Achieved evaluation metrics on training and validation data for only OV labels.

Evaluation metrics	**VGG 19**	**ResNet 152V2**	**ResNet50V2**	**MobileNetV2**	**InceptionV3**	**Inception resnet**	**Exception**	**DenseNet169**
For training data								
Overall accuracy (average of 250 epochs)	0.94	0.98	0.98	0.97	0.97	0.98	0.97	0.97
Overall accuracy (last epoch)	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98
Binary accuracy (average of 250 epochs)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Binary accuracy (last epoch)	0.99	0.99	0.99	1	0.99	0.99	0.99	0.99
Overall categorical accuracy (average of 250 epochs)	0.98	0.97	0.98	0.98	0.97	0.98	0.97	0.97
Overall categorical accuracy (last epoch)	0.98	0.97	0.98	0.98	0.98	0.99	0.98	0.98
Top-k categorical accuracy (average of 250 epochs)	1	1	1	1	1	1	1	1
Top-k categorical accuracy (last epoch)	1	1	1	1	1	1	1	1
Precision (average of 250 epochs)	0.99	0.98	0.99	0.98	0.98	0.98	0.98	0.98
Precision (last epoch)	0.99	0.99	1	1	0.99	0.99	0.99	0.99
Recall (average of 250 epochs)	0.99	0.98	0.99	0.98	0.98	0.98	0.98	0.98
Recall (last epoch)	0.99	0.99	0.99	1	0.99	0.99	0.99	0.98
Loss (average of 250 epochs)	0.09	0.25	0.09	0.07	0.08	0.08	0.06	0.08
Loss (last epoch)	0.06	0.24	0.06	0.00	0.05	0.05	0.04	0.03
For validation data								
Overall accuracy (average of 250 epochs)	0.46	0.52	0.52	0.56	0.61	0.55	0.54	0.53
Overall accuracy (last epoch)	0.47	0.52	0.52	0.55	0.60	0.54	0.55	0.52
Binary accuracy (average of 250 epochs)	0.73	0.77	0.77	0.79	0.82	0.79	0.77	0.77
Binary accuracy (last epoch)	0.70	0.78	0.78	0.77	0.83	0.78	0.77	0.77
Overall categorical accuracy (average of 250 epochs)	0.46	0.52	0.55	0.56	0.61	0.55	0.54	0.53
Overall categorical accuracy (last epoch)	0.47	0.52	0.54	0.55	0.60	0.54	0.55	0.52
Top-k categorical accuracy (average of 250 epochs)	1	1	1	1	1	1	1	1
Top-k categorical accuracy	1	1	1	1	1	1	1	1

(last epoch)								
Precision (average of 250 epochs)	0.46	0.57	0.57	0.58	0.69	0.59	0.54	0.53
Precision (last epoch)	0.43	0.57	0.56	0.54	0.74	0.59	0.55	0.53
Recall (average of 250 epochs)	0.49	0.42	0.44	0.54	0.51	0.45	0.49	0.46
Recall (last epoch)	0.63	0.42	0.45	0.56	0.47	0.46	0.48	0.46
Loss (average of 250 epochs)	1.53	5.51	5.06	6.09	3.8	2.01	3.16	3.25
Loss (last epoch)	2.84	5.52	5.03	7.31	5.8	2.25	3.89	3.68

accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss for an average of 250 epochs during the training stage was achieved up to 0.93, 0.99, 0.93, 0.99, 0.99, 0.99, and 0.04 respectively. Likewise, 0.95, 1, 0.95, 0.99, 1, 1, and 0.00 was achieved as the highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical ac- curacy, precision, recall, and loss respectively for the last epoch. The highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical ac- curacy, precision, recall, and loss for an average of 250 epochs during the validation stage was achieved up to 0.49, 0.80, 0.49, 0.90, 0.62, 0.50, and 1.28 respectively.

Table 5.7 Achieved evaluation metrics on training and validation data for both OV and VM labels.

Evaluation metrics	**VGG 19**	**ResNet152V2**	**ResNet50V2**	**MobileNetV2**	**InceptionV3**	**Inception resnet**	**Exception**	**DenseNet169**
				For training data				
Overall accuracy (average of 250 epochs)	0.52	0.93	0.90	0.82	0.84	0.88	0.80	0.76
Overall accuracy (last epoch)	0.43	0.95	0.90	0.83	0.87	0.91	0.83	0.76
Binary accuracy (average of 250 epochs)	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Binary accuracy (last epoch)	0.99	0.99	0.99	0.99	1	1	0.99	1
Overall categorical accuracy (average of 250 epochs)	0.52	0.93	0.90	0.82	0.84	0.88	0.80	0.76
Overall categorical accuracy (last epoch)	0.43	0.95	0.90	0.83	0.87	0.91	0.83	0.76
Top-k categorical accuracy (average of 250 epochs)	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Top-k categorical accuracy (last epoch)	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Precision (average of 250 epochs)	0.95	0.99	0.99	0.99	0.98	0.98	0.98	0.98
Precision (last epoch)	0.99	0.99	0.99	0.99	1	1	0.99	1
Recall (average of 250 epochs)	0.94	0.99	0.99	0.99	0.98	0.98	0.98	0.98
Recall (last epoch)	0.99	0.99	0.99	0.99	1	1	0.99	1
Loss (average of 250	0.07	0.12	0.11	0.08	0.04	0.05	0.07	0.06

epochs)								
Loss (last epoch)	0.01	0.12	0.07	0.02	0.00	0.00	0.04	0.00
For validation data								
Overall accuracy (average of 250 epochs)	0.31	0.48	0.46	0.43	0.46	0.49	0.43	0.44
Overall accuracy (last epoch)	0.28	0.49	0.48	0.44	0.48	0.51	0.46	0.44
Binary accuracy (average of 250 epochs)	0.77	0.78	0.79	0.79	0.79	0.80	0.79	0.79
Binary accuracy (last epoch)	0.76	0.78	0.78	0.80	0.79	0.80	0.78	0.79
Overall categorical accuracy (average of 250 epochs)	0.31	0.48	0.46	0.43	0.46	0.49	0.43	0.44
Overall categorical accuracy (last epoch)	0.28	0.49	0.48	0.44	0.48	0.51	0.46	0.44
Top-k categorical accuracy (average of 250 epochs)	0.88	0.79	0.89	0.89	0.87	0.85	0.84	0.90
Top-k categorical accuracy (last epoch)	0.83	0.74	0.90	0.88	0.87	0.85	0.85	0.90
Precision (average of 250 epochs)	0.54	0.58	0.62	0.60	0.60	0.62	0.60	0.59
Precision (last epoch)	0.52	0.59	0.59	0.65	0.60	0.63	0.58	0.59
Recall (average of 250 epochs)	0.48	0.41	0.43	0.47	0.48	0.50	0.49	0.50
Recall (last epoch)	0.46	0.41	0.43	0.46	0.48	0.51	0.49	0.49
Loss (average of 250 epochs)	1.28	7.40	5.41	3.24	2.79	3.01	3.85	4.00
Loss (last epoch)	1.81	8.70	5.78	2.91	2.84	3.09	5.23	4.29

Likewise, 0.51, 0.80, 0.51, 0.90, 0.65, 0.51, and 1.81 was achieved as the highest overall accuracy, binary accuracy, overall categorical accuracy, top-k categorical accuracy, precision, recall, and loss respectively for the last epoch. The highest macro average of the achieved precision, recall, and F1-score during the testing stage was achieved up to 0.31, 0.26, and 0.23 respectively. The highest weighted average of the achieved precision, re- call, and F1-score during the testing stage was achieved up to 0.69, 0.50, and 0.55 respectively. For the training and validation data, ResNet50V2, ResNet152V2, and Inception resnet performed similarly. For the test data, all algorithms performed similarly and produced poor evaluation metrics.

Table 5.8 Achieved evaluation metrics on test data for only OV, only VM, and both OV and VM labels.

Evaluation metrics	**VGG 19**	**ResNet152V2**	**ResNet50V2**	**MobileNetV2**	**InceptionV3**	**Inception resnet**	**Exception**	**DenseNet169**
(a) Only VM labels								
Precision (macro average)	0.46	0.55	0.56	0.38	0.45	0.72	0.52	0.47
Precision (weighted average)	0.53	0.60	0.59	0.54	0.56	0.71	0.60	0.54
Recall (macro average)	0.47	0.51	0.50	0.47	0.48	0.52	0.53	0.45

Recall (weighted average)	0.60	0.65	0.61	0.57	0.57	0.63	0.61	0.58
F1-score (macro average)	0.44	0.47	0.46	0.36	0.45	0.49	0.51	0.45
F1-score (weighted average)	0.54	0.58	0.53	0.52	0.55	0.60	0.59	0.55
(b) Only OV labels								
Precision (macro average)	0.46	0.52	0.55	0.56	0.51	0.52	0.54	0.55
Precision (weighted average)	0.53	0.52	0.54	0.55	0.51	0.52	0.53	0.53
Recall (macro average)	0.47	0.54	0.55	0.53	0.51	0.54	0.47	0.54
Recall (weighted average)	0.60	0.52	0.55	0.53	0.50	0.52	0.49	0.54
F1-score (macro average)	0.44	0.52	0.54	0.54	0.50	0.53	0.48	0.54
F1-score (weighted average)	0.54	0.52	0.54	0.53	0.50	0.51	0.49	0.53
(c) Both OV and VM labels								
Precision (macro average)	0.31	0.29	0.25	0.30	0.27	0.30	0.29	0.27
Precision (weighted average)	0.69	0.61	0.63	0.64	0.66	0.64	0.66	0.61
Recall (macro average)	0.23	0.16	0.19	0.26	0.17	0.20	0.18	0.18
Recall (weighted average)	0.31	0.46	0.47	0.39	0.50	0.47	0.43	0.45
F1-score (macro average)	0.13	0.19	0.21	0.16	0.20	0.23	0.21	0.21
F1-score (weighted average)	0.37	0.48	0.53	0.44	0.55	0.52	0.50	0.51

Upon application of AI on the developed dataset, a highest precision (macro average), precision (weighted average), recall (macro average), recall (micro average), F1-score (macro average), and F1-score (weighted aver- age) up-to 0.72, 0.71, 0.55, 0.65, 0.54, and 0.60 respectively was achieved on the test data for any of the three datasets. After a comprehensive evaluation of the three datasets and their AI pipeline, we noticed that several algorithms did not generalize well on the test data. A higher training evaluation metrics with slightly lower validation metrics and poor testing metrics indicated overfitting. Possible reasons include that the algorithms were not able to understand the pattern of VM and OV sub-scores, different distribution of the training, validation, and test datasets, and no hyper-parameter tuning, nor augmentation of the frames in the imbalanced dataset. The analysis also indicated that the algorithms were able to understand only VM sub-scores better in comparison to only OV sub-scores and both VM and OV sub-scores. Based on the achieved metrics, OV sub-scores were difficult to predict even with a training of 250 epochs on a super-computer. Overall, VGG19 classifier converged best over 250 epochs. Rest of the algorithms required more no. of epochs to further reach lower loss values. In future, we plan to include a real-time predictor in AI-KODA.

We show the interpretability results of each of the multi-label classification pipeline through Figure 5.3, and 5.4. Similar observations were seen through the achieved gradient activation maps. A gradient activation map offers valuable insights into a model's decision-making process. By highlighting regions of an input that heavily influence classification outcomes, it provides a spatial understanding of feature importance. High activation areas signify regions where the model concentrates its attention, typically containing critical features pertinent to the task facilitating improvements in both model design and feature engineering [122].

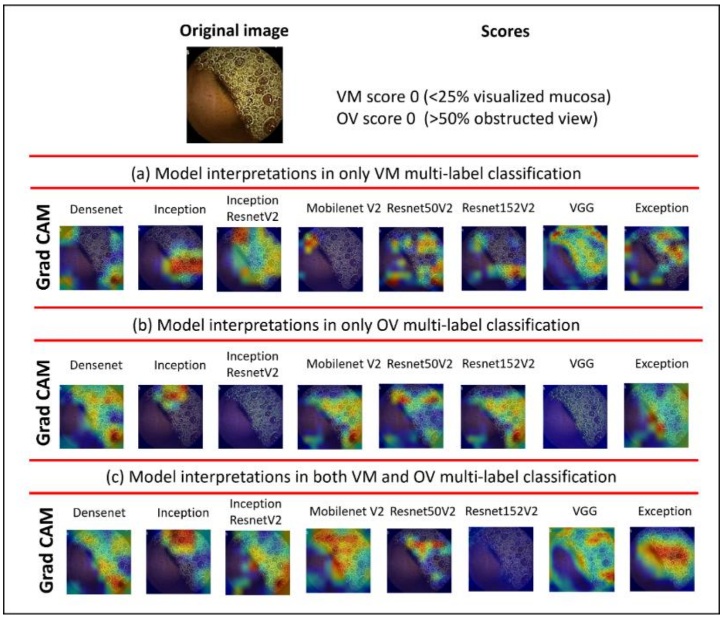

Original image
Scores
VM score 0 (<25% visualized mucosa)
OV score 0 (>50% obstructed view)
(a) Model interpretations in only VM multi-label classification
Grad CAM
Densenet
Inception
Inception ResnetV2
Mobilenet V2
Resnet50V2
Resnet152V2
VGG
Exception
(b) Model interpretations in only OV multi-label classification
(c) Model interpretations in both VM and OV multi-label classification

Figure 5.3 Gradient activation map of a VCE frame with 25% visualized mucosa and 50% obstructed view for the three multi-label classification pipelines.

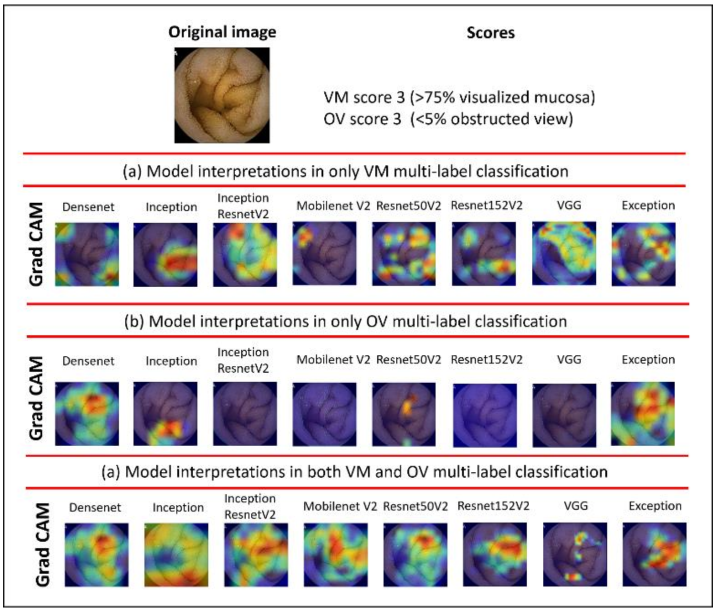

Original image
Scores
VM score 3 (>75% visualized mucosa)
OV score 3 (<5% obstructed view)
(a) Model interpretations in only VM multi-label classification
Grad CAM
Densenet
Inception
Inception ResnetV2
Mobilenet V2
Resnet50V2
Resnet152V2
VGG
Exception
(b) Model interpretations in only OV multi-label classification
(a) Model interpretations in both VM and OV multi-label classification

Figure 5.4 Gradient activation map of a VCE frame with 75% visualized mucosa and 5% obstructed view for the three multi-label classification pipelines.

Overall, the gradient activation map serves as a powerful tool for model analysis, offering actionable insights into the inner workings of machine learning algorithms and enhancing their transparency and interpretability. By visualizing these activation maps, researchers and health practitioners such as the gastroenterologists can better understand how the AI model interprets input data and make informed decisions about model architecture, training, and optimization strategies [92], [115], [116], [123], [158], [159], [160].

We discuss the strengths and limitations of the research work discussed in this chapter. One of the study's strengths is the creation of an intuitive application (AI-KODA score) that allows for real-time scoring and its mechanism to save the scores in the backend. It is fast and helps in quick assessment of the cleanliness in VCE as per KODA. In turn, the application also aids in the creation of a multi-label image collection for the automated evaluation of cleanliness in VCE.

For manual abnormality detection, around 3–4 hours are spent by experienced gastroenterologists on each video analysis of 8–12 hours of VCE [25], [28], [35], [123], [150]. During this time, the gastroenterologists also evaluate and report on the cleaning preparation and quality of VCE frames. If KODA score is implemented in a real-time clinical setting, the gastroenterologist must first choose frames at intervals of five minutes, after which they must rate each frame according to VM and OV. Analysis of every frame that was chosen will come next, and it could take a while to score each frame manually. It may require memorising all the sub-scores, or use of pen-paper to write down the scores which may lead to a high miss-rate or miss-calculation. AI-KODA score application automates both processes.

In the previous chapter, we demonstrated that the AI-KODA score application is time-saving and effective in contrast to the manual KODA score, which necessitates self-calculations and the usage of paper and pen during analysis.

Ours is the first VCE dataset to assess the cleanliness of VCE frames as per the latest medical scoring system (KODA). It is multi-labelled and of high-quality. While gathering the dataset, we did not eliminate any frames that contained abnormalities.

Abnormal frames have been excluded in previous research. This demonstrates the variety in our gathered dataset. Our study is the third to demonstrate KODA's efficacy. We concur with the original KODA study and have shown that KODA has face validity and is straightforward.

We further attempted to reduce the evaluation time with the help of AI algorithms, and automatic frame extraction to fully automate the task of cleanliness assessment in VCE. We have shown the feasibility of AI for automation in this field by applying ten machine learning algorithms and eight transfer learning algorithms on the developed dataset. KODA utilised with AI doesn't require a specific interval for collection of frames. It may potentially help in development of a generalized scoring system in VCE. The previous research has considered different types of intervals like two minutes, first and last ten minutes of the small bowel segments, first five minutes of each segment, random intervals etc while selecting the VCE frames for scoring purpose. The original KODA score considers frame selection at an interval of five minutes. In this study, two types of intervals (five minutes and random) were considered while assessing the frames as per existing KODA to check the effect of intervals with respect to its scoring system. It was observed that the ICCs of five minutes and random frames showed a similar trend in reliability estimation.

We preferred using transfer learning algorithms instead of deep learning algorithms with no changes in parameters or augmentation as they have been shown to produce good results on small biomedical datasets[39], [127], [128], [129], [130], [150]. We have reported all the common evaluations used in multi-label classification over a 250-epoch training with distinct training, validation, and test datasets while ensuring no data leakage and repetition of frames for the transfer learning algorithms. A similar methodology was followed for the machine learning algorithms based multi-label classification. Through this, we intended to help the future researchers to compare their work with our standardised results and improve this state-of-the-art study.

There are a few limitations to the present study discussed. First, video level analysis was not performed as the original KODA doesn’t have specific instructions for the same. In this research field, there is a mixed opinion for use of frames v/s a full-length video for automatic cleanliness assessment. As per the position statement of ESGE on the expected value of AI [23], automatic assessment of cleanliness in VCE should be evaluated for full length video as per existing validated scales. However, presently, frame-by-frame analysis is more preferred in clinical practice as the full-length video is about $8 - 12$ hours [23]. The full-length videos are not directly downloadable and can only be viewed on the CE soft wares like RAPID. Our system allows a video of 20 seconds for downloading purpose. Second, AI-KODA score application is only available for WCE users now. The users must login from the same Android device while scoring the frames. Improvements are being done on this limitation. History information was not saved to maintain the confidentiality of the frames and the users. Third, the collected data is from one centre and contains third generation CE frames. We are in process of collecting scores of the open-source datasets in VCE like the Kvasir-capsule and CE-cleanliness. Despite these limitations, the present study has achieved encouraging results.

5.6 Conclusion and Future Scope

The use of automatic cleanliness assessment in VCE is essential as it has the potential to mitigate the subjectivity, complexity, and duration of the current scoring systems in this research field. In this study, a new automatic scoring system for evaluating the cleanliness of VCE frames was developed. Initially, a novel multi-label image dataset comprising medical scores of twenty-eight VCE frames was produced via the suggested mobile ap- plication named AI-KODA score. Second, with the assistance of three gastroenterologists, the efficacy of the collected data was examined. Third, to demonstrate the potential of AI in this research field, a comprehensive evaluation and benchmarking of the created dataset was carried out with the assistance of ten machine learning and eight transfer learning algorithms. With positive outcomes, the study is simple to apply in both clinical and research settings.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

In this thesis, we addressed two tasks namely automatic detection of abnormality in polyp and non-polyp frame in colonoscopy frames and automatic assessment of cleanliness in VCE. The proposed methodologies for these two tasks are scalable, robust, work in real-time, and are explainable in nature. The comprehensive analysis followed for each of the task, shows its promising future in the gastroenterology department. We discuss the summary of the work done in the thesis through Section 6.1. It is followed by future scope/ directions and social impact of the work done in Section 6.2.

6.1 Summary of the Work Done in the Thesis

The first task focused on developing an explainable, end-to-end WADT-MCPI architecture for automatic colorectal polyp diagnosis using colonoscopy CP and non-CP frames. The developed architecture consisted of a novel, fine-tuned feature-extracting module, followed by CP and non-CP frame identification and a window-based CP detection system. The architecture achieved an overall accuracy, precision, recall, specificity, F1 score, and AUC score up to 94.23%, 91.16%, 94.00%, 92.67%, 91.75%, and 92.53% respectively on the colonoscopy dataset. To show the robustness of the developed architecture, a new test set was developed and evaluated. After the analysis, it released on Zenodo, an open-source platform for research purposes. It is called the gastrointestinal atlas-colon polyp dataset. It contains seven patient videos obtained from open-source, copyright free web sources. Explainable and evaluation methods like class activation mapping, feature mapping, occlusion testing, hyper-parameter tuning ablation experiments, and separate, sequential, and non-sequential frame-based test set analysis were used to show the efficacy of the proposed architecture. The developed architecture was compared with the existing state-of-the-art methodologies in this field. Additionally, the architecture was compared with a transfer learning architecture as well. In both the comparisons, the developed architecture outperformed and achieved better evaluation metrics.

The second task focused on development of methodology to automatically assess the cleanliness in VCE video frames. First, we automated the process of scoring VCE video frames as per existing KODA scoring system by developing an easy-to-use mobile application called AI-KODA. AI-KODA Score is a flutter-based application which can be downloaded on a mobile. The application first trains a gastroenterologist how to use KODA. After a simple training, the gastroenterologist can upload VCE video frames on the application and score them. After successful scoring, a report is generated for the overall score. The scores are also collected in real-time and saved for the development of a frame level, high-quality, and multi-labelled dataset for automatic multi-label classification of clean v/s dirty VCE video frames. The developed dataset was subjective to medical verification with the help of three experienced gastroenterologists. A good level of ICCs was achieved. Bases on the common consensus by the three gastroenterologists, a common dataset comprising of 2173 with eight distinct labels of KODA were developed. A comprehensive evaluation, interpretation, benchmarking of the generated dataset was done using ten machine learning models and eight transfer learning algorithms on google Collaboratory and a supercomputer named, NVIDIA RTX A5000 workstation. The developed dataset and its methodology are first-of-its-kind. With positive outcomes, the study is simple to apply in both clinical and research settings and helps in automatic assessment of cleanliness in VCE. Both the methodologies developed in the first and second task helps in reducing the time and effort of the gastroenterologist in timely detection of polyps in colonoscopy and cleanliness assessment in VCE.

6.2 Future Scope/ Directions and Social Impact

As we peer into the horizon of possibilities of automation in the field of polyp detection in colonoscopy and cleanliness assessment in VCE through AI, we are propelled by the conviction that the pursuit of knowledge and research is an ever-evolving process, one that demands continuous exploration, adaptation, and refinement. We discuss the possible future directions and social impact in these fields.

For the first task, the future scope may in the direction to a large, multi-centre clinical study to validate the developed architecture which may also:

- Improve the evaluation metrics.
- Inclusion of more colonoscopy data with different types of occlusion effects.
- Focus on selection bias problem in this field.
- Comprehensive evaluation using more advanced explainability techniques.

For the second task, the future scope may be in the direction to a large, multi-centre clinical study to validate KODA score which may also:

- Determine an acceptable cut-off ratio.
- Develop a method for video analysis.
- Improve the intra-rater reliability estimates.
- Further simplify OV and add more training examples.
- Comprehensive evaluation of the developed dataset with meta-learning algorithms.
- Development of multi-head algorithms for combined and robust prediction of VM and OV.

REFERENCES

[1] L. R. Johnson, *Physiology of the gastrointestinal tract*. Elsevier, 2006.

[2] A. Gross, “Gastrointestinal Diseases Rise in Asia.” [Online]. Available: https://www.medtechintelligence.com/feature_article/gastrointestinal-diseases-rise-in-asia/

[3] D. Shah *et al.*, “Burden of Gastrointestinal and Liver Diseases in India, 1990–2016,” *Indian Journal of Gastroenterology*, vol. 37, no. 5, pp. 439–445, 2018.

[4] “Burden of Digestive Diseases - PAHO/WHO | Pan American Health Organization.” Accessed: Mar. 20, 2023. [Online]. Available: https://www.paho.org/en/noncommunicable-diseases-and-mental-health/noncommunicable-diseases-and-mental-health-data-38

[5] R. Wang, Z. Li, and D. Zhang, “Global, Regional, And National Burden Of 10 Digestive Diseases In 204 Countries and Territories From 1990 To 2019,” *Front Public Health*, vol. 11, p. 1061453, 2023.

[6] M. Arnold *et al.*, “Global Burden of 5 Major Types of Gastrointestinal Cancer,” *Gastroenterology*, vol. 159, no. 1, pp. 335–349, 2020.

[7] M. Hartman *et al.*, “National Health Care Spending In 2018: Growth Driven by Accelerations In Medicare And Private Insurance Spending: US health care spending increased 4.6 percent to reach $3.6 trillion in 2018, a faster growth rate than that of 4.2 percent in 2017 but the same rate as in 2016.,” *Health Aff*, vol. 39, no. 1, pp. 8–17, 2020.

[8] “8.4% CAGR for Capsule Endoscopy Market Size Worth $757.51.” Accessed: Mar. 20, 2023. [Online]. Available: https://www.globenewswire.com/en/news-release/2023/03/09/2624174/0/en/8-4-CAGR-for-Capsule-Endoscopy-Market-Size-Worth-757-51-Million-by-2028-Deep-Dive-Analysis-of-18-Countries-across-5-Key-Regions-50-Companies-Scrutinized-The-Insight-Partners.html

[9] H. Shinya and W. Wolff, “Flexible Colonoscopy,” *Cancer*, vol. 37, no. S1, pp. 462–470, 1976.

[10] D. K. Rex *et al.*, “Quality Indicators for Colonoscopy,” *Gastrointest Endosc*, vol. 63, no. 4, pp. S16–S28, 2006.

[11] R. Niikura *et al.*, “Colonoscopy Reduces Colorectal Cancer Mortality: A

Multicenter, Long-Term, Colonoscopy-Based Cohort Study," *PLoS One*, vol. 12, no. 9, p. e0185294, 2017.

[12] N. N. Baxter *et al.*, "Association of Colonoscopy and Death from Colorectal Cancer," *Ann Intern Med*, vol. 150, no. 1, pp. 1–8, 2009.

[13] M. B. Nierengarten, "Colonoscopy Remains the Gold Standard for Screening Despite Recent Tarnish: Although A Recent Study Seemed to Indicate That Colonoscopies Are Not as Effective As Once Thought At Detecting Colorectal Cancer, A Closer Look At The Study Clears The Confusion." Wiley Online Library, 2023.

[14] G. Iddan *et al.*, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, p. 417, 2000.

[15] R. Gupta, "The Journey of Capsule Endoscopy in India," *Journal of Digestive Endoscopy*, vol. 9, no. 04, p. 183, 2018.

[16] S. M. Khanna *et al.*, "Therapeutic Endoscopy-Review of Current Guidelines and Recommendations-IAGES This document was circulated, reviewed and approved by the Executive council of IAGES in the EC Meeting held on Shillong 14 th September 2018 at Shillong, India. IAGES Therapeutic Endoscopy Guidelines Committee: -Committee Members," 2018.

[17] J. Oh *et al.*, "Measuring Objective Quality of Colonoscopy," *IEEE Trans Biomed Eng*, vol. 56, no. 9, pp. 2190–2196, 2008.

[18] V. Jahmunah *et al.*, "Endoscopy, Video Capsule Endoscopy, And Biopsy for Automated Celiac Disease Detection: A Review," *Biocybern Biomed Eng*, vol. 43, no. 1, pp. 82–108, 2023.

[19] G. Ciuti, A. Menciassi, and P. Dario, "Capsule Endoscopy: From Current Achievements to Open Challenges," *IEEE Rev Biomed Eng*, vol. 4, pp. 59–72, 2011.

[20] M. Keuchel *et al.*, "Quantitative Measurements in Capsule Endoscopy," *Comput Biol Med*, vol. 65, pp. 333–347, 2015.

[21] L. Y. Korman *et al.*, "Capsule Endoscopy Structured Terminology (CEST): Proposal of A Standardized and Structured Terminology for Reporting Capsule Endoscopy Procedures," *Endoscopy*, vol. 37, no. 10, pp. 951–959, 2005.

[22] K.-N. Shim *et al.*, "Quality Indicators for Small Bowel Capsule Endoscopy,"

Clin Endosc, vol. 50, no. 2, p. 148, 2017.

[23] E. Rondonotti *et al.*, "Quality Performance Measures for Small Capsule Endoscopy: Are The ESGE Quality Standards Met?" *Endosc Int Open*, vol. 9, no. 02, pp. E122–E129, 2021.

[24] K. Muhammad *et al.*, "Vision-Based Personalized Wireless Capsule Endoscopy For Smart Healthcare: Taxonomy, Literature Review, Opportunities And Challenges," *Future Generation Computer Systems*, vol. 113, pp. 266–280, 2020.

[25] A. Ponte *et al.*, "Review Of Small-Bowel Cleansing Scales In Capsule Endoscopy: A Panoply Of Choices," *World J Gastrointest Endosc*, vol. 8, no. 17, p. 600, 2016.

[26] C. Sánchez-Montes *et al.*, "Review of Computational Methods for The Detection And Classification Of Polyps In Colonoscopy Imaging," *Gastroenterología y Hepatología (English Edition)*, vol. 43, no. 4, pp. 222–232, 2020.

[27] L. Houwen *et al.*, "Comprehensive Review of Publicly Available Colonoscopic Imaging Databases for Artificial Intelligence Research: Availability, Accessibility, And Usability," *Gastrointest Endosc*, vol. 97, no. 2, pp. 184–199, 2023.

[28] P. Handa, N. Goel, and S. Indu, "Datasets of Wireless Capsule Endoscopy For AI-Enabled Techniques," in *International Conference on Computer Vision and Image Processing*, Springer, 2021, pp. 439–446.

[29] L. F. Sanchez-Peralta *et al.*, "Deep Learning to Find Colorectal Polyps in Colonoscopy: A Systematic Literature Review," *Artif Intell Med*, vol. 108, p. 101923, 2020.

[30] W.-N. Liu *et al.*, "Study on Detection Rate of Polyps and Adenomas In Artificial-Intelligence-Aided Colonoscopy," *Saudi J Gastroenterol*, vol. 26, no. 1, p. 13, 2020.

[31] Y. Mori *et al.*, "Cost Savings in Colonoscopy with Artificial Intelligence-Aided Polyp Diagnosis: An Add-On Analysis of A Clinical Trial (With Video)," *Gastrointest Endosc*, vol. 92, no. 4, pp. 905–911, 2020.

[32] F. J. O'Hara and D. Mc Namara, "Capsule Endoscopy with Artificial

Intelligence-Assisted Technology: Real-World Usage of a Validated AI Model For Capsule Image Review,” *Endosc Int Open*, vol. 11, no. 10, pp. E970–E975, 2023.

[33] A. P. Abadir *et al.*, “Artificial Intelligence in Gastrointestinal Endoscopy,” *Clin Endosc*, vol. 53, no. 2, p. 132, 2020.

[34] G. Antonelli *et al.*, “Impact of Artificial Intelligence on Colorectal Polyp Detection,” *Best Pract Res Clin Gastroenterol*, vol. 52, p. 101713, 2021.

[35] D. Gunjan *et al.*, “Small Bowel Bleeding: A Comprehensive Review,” *Gastroenterol Rep (Oxf)*, vol. 2, no. 4, pp. 262–275, 2014.

[36] A. P. Abadir *et al.*, “Artificial Intelligence in Gastrointestinal Endoscopy,” *Clin Endosc*, vol. 53, no. 2, p. 132, 2020.

[37] J. Mongan, L. Moy, and C. E. Kahn Jr, “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers,” *Radiol Artif Intell*, vol. 2, no. 2, 2020.

[38] Y. J. Yang, “The Future of Capsule Endoscopy: The Role of Artificial Intelligence and Other Technical Advancements,” *Clin Endosc*, vol. 53, no. 4, p. 387, 2020.

[39] P. Kora *et al.*, “Transfer Learning Techniques for Medical Image Analysis: A Review,” *Biocybern Biomed Eng*, vol. 42, no. 1, pp. 79–107, 2022.

[40] H. Mangotra, P. Handa, and N. Goel, “Open-source Datasets for Colonoscopy Polyps and its AI-enabled Techniques,” in *Communication and Intelligent Systems - Proceedings of ICCIS 2022, Volume 1*, H. Sharma, V. Shrivastava, K. K. Bharti, and L. Wang, Eds., Springer, 2022.

[41] J. Bernal, J. Sánchez, and F. Vilarino, “Towards Automatic Polyp Detection with A Polyp Appearance Model,” *Pattern Recognit*, vol. 45, no. 9, pp. 3166–3182, 2012.

[42] J. Silva *et al.*, “Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer,” *Int J Comput Assist Radiol Surg*, vol. 9, pp. 283–293, 2014.

[43] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information,” *IEEE Trans Med Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[44] P. Mesejo *et al.*, “Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy,” *IEEE Trans Med Imaging*, vol. 35, no. 9, pp. 2051–2063, 2016.

[45] J. Bernal *et al.*, “Wm-Dova Maps for Accurate Polyp Highlighting in Colonoscopy: Validation Vs. Saliency Maps from Physicians,” *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[46] J. Bernal *et al.*, “Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from The Miccai 2015 Endoscopic Vision Challenge,” *IEEE Trans Med Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[47] D. Vázquez *et al.*, “A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images,” *J Healthc Eng*, vol. 2017.

[48] G.-P. Ji *et al.*, “Progressively Normalized Self-Attention Network for Video Polyp Segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 142–152.

[49] K. Pogorelov *et al.*, “Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 164–169.

[50] H. Borgli *et al.*, “Hyperkvasir, A Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy,” *Sci Data*, vol. 7, no. 1, p. 283, 2020.

[51] D. Jha *et al.*, “Kvasir-Seg: A Segmented Polyp Dataset,” in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, Springer, 2020, pp. 451–462.

[52] L. F. Sánchez-Peralta *et al.*, “Piccolo White-Light and Narrow-Band Imaging Colonoscopic Dataset: A Performance Comparative of Models and Datasets,” *Applied Sciences*, vol. 10, no. 23, p. 8501, 2020.

[53] S. Ali *et al.*, “Polypgen: A Multi-Center Polyp Detection and Segmentation Dataset for Generalisability Assessment,” *arXiv preprint arXiv:2106.04463*, 2021.

[54] W. Wang *et al.*, “An Improved Deep Learning Approach and Its Applications on Colonic Polyp Images Detection,” *BMC Med Imaging*, vol. 20, pp. 1–14, 2020.

[55] M. Misawa *et al.*, “Development of A Computer-Aided Detection System for

Colonoscopy and A Publicly Accessible Large Colonoscopy Video Database (With Video),” *Gastrointest Endosc*, vol. 93, no. 4, pp. 960–967, 2021.

[56] Y. Ma *et al.*, “Ldpolypvideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, Springer, 2021, pp. 387–396.

[57] P. Ngoc Lan *et al.*, “Neounet: Towards Accurate Colon Polyp Segmentation and Neoplasm Detection,” in *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*, Springer, 2021, pp. 15–28.

[58] D. Fitting *et al.*, “A Video Based Benchmark Data Set (ENDOTEST) To Evaluate Computer-Aided Polyp Detection Systems,” *Scand J Gastroenterol*, vol. 57, no. 11, pp. 1397–1403, 2022.

[59] K. Li *et al.*, “Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations,” *PLoS One*, vol. 16, no. 8, p. e0255809, 2021.

[60] N. S. An *et al.*, “Blazeneo: Blazing Fast Polyp Segmentation and Neoplasm Detection,” *IEEE Access*, vol. 10, pp. 43669–43684, 2022.

[61] K. Li *et al.*, “Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations,” *PLoS One*, vol. 16, no. 8, p. e0255809, 2021.

[62] A. Goyal *et al.*, “Automatic Detection of WCE Bleeding Frames Using Hybrid Features and Machine Learning Algorithms,” in *2022 IEEE India Council International Subsections Conference (INDISCON)*, IEEE, 2022, pp. 1–7.

[63] P. Handa *et al.*, “Auto-WCEBleedGen Version V1 and V2: Challenge, Datasets and Evaluation,” *Authorea Preprints*, 2024.

[64] P. Handa, N. Goel, and S. Indu, “Datasets of Wireless Capsule Endoscopy for AI-Enabled Techniques,” in *Computer Vision and Image Processing*, B. Raman, S. Murala, A. Chowdhury, A. Dhall, and P. Goyal, Eds., Cham: Springer International Publishing, 2022, pp. 439–446.

[65] P. H. Smedsrud *et al.*, “Kvasir-Capsule, A Video Capsule Endoscopy Dataset,” *Sci Data*, vol. 8, no. 1, pp. 1–10, 2021.

[66] M. Haslach-Häfner and K. Mönkemüller, “Reading Capsule Endoscopy: Why

Not AI Alone?” *Endosc Int Open*, vol. 11, no. 12, pp. E1175–E1176, 2023.

[67] V. Vats, P. Goel, A. Agarwal, and N. Goel, “SURF-SVM Based Identification and Classification of Gastrointestinal Diseases in Wireless Capsule Endoscopy,” *arXiv preprint arXiv:2009.01179*, 2020.

[68] A. Koulaouzidis *et al.*, “KID Project: An Internet-Based Digital Video Atlas of Capsule Endoscopy for Research Purposes,” *Endosc Int Open*, vol. 5, no. 06, pp. E477–E483, 2017.

[69] “The Gastrolab Image Gallery.” Accessed: Jun. 15, 2022. [Online]. Available: https://gastrolab.net/gall258.htm.

[70] F. Deeba, F. M. Bui, and K. A. Wahid, “Automated Grow-cut For Segmentation of Endoscopic Images,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 4650–4657. [Online]. Available: https://sites.google.com/site/farahdeeba073/Research/resources

[71] R. Vallée, A. De Maissin *et al.*, “Crohnipi: An Endoscopic Image Database for The Evaluation of Automatic Crohn’s Disease Lesions Recognition Algorithms,” in *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, International Society for Optics and Photonics, 2020, p. 113171Q.

[72] K. B. Ozyoruk *et al.*, “Endoslam Dataset and An Unsupervised Monocular Visual Odometry and Depth Estimation Approach for Endoscopic Videos,” *Med Image Anal*, vol. 71, p. 102058, 2021.

[73] K. İncetan *et al.*, “VR-Caps: A Virtual Environment for Capsule Endoscopy,” *Med Image Anal*, vol. 70, p. 101990, 2021.

[74] R. Noorda *et al.*, “Automatic Evaluation of Degree of Cleanliness in Capsule Endoscopy Based on A Novel CNN Architecture,” *Sci Rep*, vol. 10, no. 1, p. 17706, 2020.

[75] L. F. Sánchez-Peralta *et al.*, “Deep Learning to Find Colorectal Polyps in Colonoscopy: A Systematic Literature Review,” *Artif Intell Med*, vol. 108, p. 101923, 2020.

[76] K. Mohandas, “Colorectal Cancer in India: Controversies, Enigmas and Primary Prevention,” *Indian Journal of Gastroenterology*, vol. 30. Springer, pp. 3–6, 2011.

[77] B. Bressler *et al.*, “Rates of New or Missed Colorectal Cancers After Colonoscopy and Their Risk Factors: A Population-Based Analysis,” *Gastroenterology*, vol. 132, no. 1, pp. 96–102, 2007.

[78] M. B. Nierengarten, “Colonoscopy Remains the Gold Standard for Screening Despite Recent Tarnish: Although A Recent Study Seemed to Indicate That Colonoscopies Are Not as Effective as Once Thought at Detecting Colorectal Cancer, A Closer Look At The Study Clears The Confusion.” Wiley Online Library, 2023.

[79] D. K. Iakovidis *et al.*, “Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification,” *IEEE Trans Med Imaging*, vol. 37, no. 10, pp. 2196–2210, 2018.

[80] S. Jain *et al.*, “A Deep CNN Model for Anomaly Detection and Localization in Wireless Capsule Endoscopy Images,” *Comput Biol Med*, vol. 137, p. 104789, 2021.

[81] P. Lafeuille *et al.*, “Flat Colorectal Sessile Serrated Polyp: An Example of What Artificial Intelligence Does Not Easily Detect,” *Endoscopy*, vol. 54, no. 05, pp. 520–521, 2022.

[82] Cristian R Munteanu, “Colonoscopy Polyps Detection with CNNs.” [Online]. Available: https://github.com/muntisa/Colonoscopy-polyps-detection-with-CNNs

[83] S. J. Stryker *et al.*, “Natural History of Untreated Colonic Polyps,” *Gastroenterology*, vol. 93, no. 5, pp. 1009–1013, 1987.

[84] D.-C. Cheng *et al.*, “Colorectal Polyps Detection Using Texture Features and Support Vector Machine,” in *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry: Third International Conference, MDA 2008 Leipzig, Germany, July 14, 2008, Proceedings 3*, Springer, 2008, pp. 62–72.

[85] S. Ameling *et al.*, “Texture-Based Polyp Detection in Colonoscopy,” in *Bildverarbeitung für die Medizin 2009: Algorithmen—Systeme—Anwendungen Proceedings des Workshops vom 22. bis 25. März 2009 in Heidelberg*, Springer, 2009, pp. 346–350.

[86] S. Hwang, “Automatic content analysis of endoscopy video (Endoscopic

multimedia information system.” Doctoral Thesis. The University of Texas at Arlington, 2007.

[87] L. Ruiz *et al.*, “COLON: The Largest Colonoscopy Long Sequence Public Database,” *arXiv preprint arXiv:2403.00663*, 2024.

[88] D.-C. Cheng *et al.*, “Automatic Detection of Colorectal Polyps in Static Images,” *Biomed Eng (Singapore)*, vol. 23, no. 05, pp. 357–367, 2011.

[89] L. Kliegis *et al.*, “Can A Polyp Detection and Characterization System Predict Complete Resection?” *Digestive Diseases*, vol. 40, no. 1, pp. 115–118, 2022.

[90] W.-L. Chao, H. Manickavasagan, and S. G. Krishna, “Application of Artificial Intelligence in The Detection and Differentiation Of Colon Polyps: A Technical Review For Physicians,” *Diagnostics*, vol. 9, no. 3, p. 99, 2019.

[91] S. Shah *et al.*, “Effect of Computer-Aided Colonoscopy on Adenoma Miss Rates and Polyp Detection: A Systematic Review and Meta-Analysis,” *J Gastroenterol Hepatol*, vol. 38, no. 2, pp. 162–176, 2023.

[92] S. Wang *et al.*, “An Interpretable Deep Neural Network for Colorectal Polyp Diagnosis Under Colonoscopy,” *Knowl Based Syst*, vol. 234, p. 107568, 2021.

[93] X. Zhang *et al.*, “Real-Time Gastric Polyp Detection Using Convolutional Neural Networks,” *PLoS One*, vol. 14, no. 3, p. e0214133, 2019.

[94] C.-M. Hsu *et al.*, “Colorectal Polyp Image Detection and Classification Through Grayscale Images and Deep Learning,” *Sensors*, vol. 21, no. 18, p. 5995, 2021.

[95] E. Soons *et al.*, “Real-Time Colorectal Polyp Detection Using a Novel Computer-Aided Detection System (Cade): A Feasibility Study,” *Int J Colorectal Dis*, vol. 37, no. 10, pp. 2219–2228, 2022.

[96] D. Jha *et al.*, “Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning,” *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[97] H. A. Qadir *et al.*, “Improving Automatic Polyp Detection Using CNN By Exploiting Temporal Dependency in Colonoscopy Video,” *IEEE J Biomed Health Inform*, vol. 24, no. 1, pp. 180–193, 2019.

[98] A. Nogueira-Rodríguez *et al.*, “Real-Time Polyp Detection Model Using Convolutional Neural Networks,” *Neural Comput Appl*, vol. 34, no. 13, pp.

10375–10396, 2022.

[99] F. Nehme *et al.*, “Performance and Attitudes Toward Real-Time Computer-Aided Polyp Detection During Colonoscopy in A Large Tertiary Referral Center in The United States,” *Gastrointest Endosc*, 2023.

[100] A. Krenzer *et al.*, “A Real-Time Polyp-Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks,” *J Imaging*, vol. 9, no. 2, p. 26, 2023.

[101] A. Klein *et al.*, “Validated Computed Cleansing Score for Video Capsule Endoscopy,” *Digestive Endoscopy*, vol. 28, no. 5, pp. 564–569, 2016.

[102] O. Pietri *et al.*, “Development and Validation of An Automated Algorithm to Evaluate the Abundance of Bubbles in Small Bowel Capsule Endoscopy,” *Endosc Int Open*, vol. 6, no. 04, pp. E462–E469, 2018.

[103] E. Abou Ali *et al.*, “Development and Validation of a Computed Assessment of Cleansing Score for Evaluation of Quality Of Small-Bowel Visualization In Capsule Endoscopy,” *Endosc Int Open*, vol. 6, no. 06, pp. E646–E651, 2018.

[104] S. Oumrani *et al.*, “Multi-Criterion, Automated, High-Performance, Rapid Tool for Assessing Mucosal Visualization Quality of Still Images in Small Bowel Capsule Endoscopy,” *Endosc Int Open*, vol. 7, no. 08, pp. E944–E948, 2019.

[105] J. H. Nam *et al.*, “Development of A Deep Learning-Based Software for Calculating Cleansing Score in Small Bowel Capsule Endoscopy,” *Sci Rep*, vol. 11, no. 1, pp. 1–8, 2021.

[106] R. Leenhardt *et al.*, “A Neural Network-Based Algorithm for Assessing the Cleanliness of Small Bowel During Capsule Endoscopy,” *Endoscopy*, vol. 53, no. 09, pp. 932–936, 2021.

[107] M. D. Vasilakakis *et al.*, “Weakly Supervised Multilabel Classification for Semantic Interpretation of Endoscopy Video Frames,” *Evolving Systems*, vol. 11, no. 3, pp. 409–421, 2020.

[108] Y.-S. Park and J.-W. Lee, “Class-Labeling Method for Designing a Deep Neural Network Of Capsule Endoscopic Images Using A Lesion-Focused Knowledge Model,” *Journal of Information Processing Systems*, vol. 16, no. 1, pp. 171–183, 2020.

[109] A. Mohammed *et al.*, “PS-DeVCEM: Pathology-Sensitive Deep Learning

Model for Video Capsule Endoscopy Based on Weakly Labeled Data," *Computer Vision and Image Understanding*, vol. 201, p. 103062, 2020.

[110] M. Y. Chan, H. Cohen, and B. M. R. Spiegel, "Fewer Polyps Detected by Colonoscopy as The Day Progresses at A Veteran's Administration Teaching Hospital," *Clinical Gastroenterology and Hepatology*, vol. 7, no. 11, pp. 1217–1223, 2009.

[111] N. Goel *et al.*, "Investigating the Significance of Color Space for Abnormality Detection in Wireless Capsule Endoscopy Images," *Biomed Signal Process Control*, vol. 75, p. 103624, 2022.

[112] A. Nogueira-Rodríguez *et al.*, "Deep Neural Networks Approach for Detecting and Classifying Colorectal Polyps," *Neurocomputing*, vol. 423, pp. 721–734, 2021.

[113] N. Goel *et al.*, "Dilated CNN For Abnormality Detection in Wireless Capsule Endoscopy Images," *Soft comput*, pp. 1–17, 2022.

[114] T. Rahim, S. A. Hassan, and S. Y. Shin, "A Deep Convolutional Neural Network for The Detection of Polyps in Colonoscopy Images," *Biomed Signal Process Control*, vol. 68, p. 102654, 2021.

[115] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and Interpretability in Convolutional Neural Networks For Semantic Segmentation Of Colorectal Polyps," *Med Image Anal*, vol. 60, p. 101619, 2020.

[116] H. Mangotra, P. Handa and N. Goel, "Effect of Selection Bias on Automatic Colonoscopy Polyp Detection," *Biomed Signal Process Control*, vol. 85, p. 104915, 2023.

[117] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[118] A. F. Agarap, "Deep Learning Using Rectified Linear Units (Relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[119] P. Lakhani *et al.*, "Hello World Deep Learning in Medical Imaging," *J Digit Imaging*, vol. 31, pp. 283–289, 2018.

[120] V. Dumoulin and F. Visin, "A Guide to Convolution Arithmetic for Deep Learning," *arXiv preprint arXiv:1603.07285*, 2016.

[121] K. Kavukcuoglu *et al.*, "Learning Convolutional Feature Hierarchies for Visual Recognition," *Adv Neural Inf Process Syst*, vol. 23, 2010.

[122] R. R. Selvaraju *et al.*, “Grad-CAM: Visual Explanations from Deep Networks Via Gradient-Based Localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[123] T. Singh, P. Handa, and N. Goel, “Automatic GI Bleeding Detection: A Comparative Analysis of Pre-Trained Deep Learning Architectures,” in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, 2023, pp. 260–265.

[124] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A Deeper Look at Dataset Bias,” in *Domain adaptation in computer vision applications*, Springer, 2017, pp. 37–55.

[125] E. J. Hegedus and J. Moody, “Clinimetrics Corner: The Many Faces of Selection Bias,” *Journal of Manual & Manipulative Therapy*, vol. 18, no. 2, pp. 69–73, 2010.

[126] H. Mangotra, *et al.*, “Effect of Selection Bias on Automatic Colonoscopy Polyp Detection,” *Biomed Signal Process Control*, vol. 85, p. 104915, 2023.

[127] Fchollet, “Transfer learning & fine-tuning.” Accessed: Oct. 11, 2022. [Online]. Available: https://keras.io/guides/transfer_learning/

[128] M. Romero *et al.*, “Targeted Transfer Learning to Improve Performance in Small Medical Physics Datasets,” *Med Phys*, vol. 47, no. 12, pp. 6246–6256, 2020.

[129] L. Alzubaidi *et al.*, “Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data,” *Cancers (Basel)*, vol. 13, no. 7, p. 1590, 2021.

[130] L. Du, “How Much Deep Learning Does Neural Style Transfer Really Need? An Ablation Study,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3150–3159.

[131] K. Patel *et al.*, “A Comparative Study on Polyp Classification Using Convolutional Neural Networks,” *PLoS One*, vol. 15, no. 7, p. e0236452, 2020.

[132] Y.-C. Jheng *et al.*, “A Novel Machine Learning-Based Algorithm to Identify and Classify Lesions and Anatomical Landmarks In Colonoscopy Images,” *Surg Endosc*, vol. 36, pp. 640–650, 2022.

[133] X. Jia *et al.*, “Automatic Polyp Recognition in Colonoscopy Images Using Deep Learning and Two-Stage Pyramidal Feature Prediction,” *IEEE Transactions on*

Automation Science and Engineering, vol. 17, no. 3, pp. 1570–1584, 2020.

[134] A. Ellahyani *et al.*, “Fine-Tuned Deep Neural Networks for Polyp Detection in Colonoscopy Images,” *Pers Ubiquitous Comput*, vol. 27, no. 2, pp. 235–247, 2023.

[135] M. K. Goenka, S. Majumder, and U. Goenka, “Capsule Endoscopy: Present Status and Future Expectation,” *World Journal of Gastroenterology: WJG*, vol. 20, no. 29, p. 10024, 2014.

[136] R. Gupta, “The Journey of Capsule Endoscopy in India,” *Journal of Digestive Endoscopy*, vol. 9, no. 04, p. 183, 2018.

[137] B. Rosa *et al.*, “Scoring Systems in Clinical Small-Bowel Capsule Endoscopy: All You Need to Know!” *Endosc Int Open*, vol. 9, no. 06, pp. E802–E823, 2021.

[138] N. Viazis *et al.*, “Bowel Preparation Increases the Diagnostic Yield of Capsule Endoscopy: A Prospective, Randomized, Controlled Study,” *Gastrointest Endosc*, vol. 60, no. 4, pp. 534–538, 2004.

[139] C. Brotz *et al.*, “A Validation Study Of 3 Grading Systems to Evaluate Small-Bowel Cleansing for Wireless Capsule Endoscopy: A Quantitative Index, A Qualitative Evaluation, And an Overall Adequacy Assessment,” *Gastrointest Endosc*, vol. 69, no. 2, pp. 262–270, 2009.

[140] M. Alageeli *et al.*, “KODA Score: An Updated and Validated Bowel Preparation Scale for Patients Undergoing Small Bowel Capsule Endoscopy,” *Endosc Int Open*, vol. 8, no. 08, pp. E1011–E1017, 2020.

[141] S. C. Park *et al.*, “A Novel Cleansing Score System for Capsule Endoscopy,” *World journal of gastroenterology: WJG*, vol. 16, no. 7, p. 875, 2010.

[142] M. Sey *et al.*, “A Randomized Controlled Trial of High-Volume Simethicone to Improve Visualization During Capsule Endoscopy,” *PLoS One*, vol. 16, no. 4, p. e0249490, 2021.

[143] S. D. Walter, M. Eliasziw, and A. Donner, “Sample Size and Optimal Designs for Reliability Studies,” *Stat Med*, vol. 17, no. 1, pp. 101–110, 1998.

[144] K. A. Batterton and K. N. Hale, “The Likert Scale What It Is and How to Use It,” *Phalanx*, vol. 50, no. 2, pp. 32–39, 2017, [Online]. Available: http://www.jstor.org/stable/26296382

[145] J. Ju *et al.*, “Clean Mucosal Area Detection of Gastroenterologists Versus

Artificial Intelligence in Small Bowel Capsule Endoscopy," *Medicine*, vol. 102, no. 6, 2023.

[146] N. Goel *et al.*, "Dilated CNN For Abnormality Detection in Wireless Capsule Endoscopy Images," *Soft comput*, pp. 1–17, 2022.

[147] J. Liu and X. Yuan, "Obscure Bleeding Detection in Endoscopy Images Using Support Vector Machines," *Optimization and engineering*, vol. 10, no. 2, pp. 289–299, 2009.

[148] C. Hong-Bin *et al.*, "Evaluation of Visualized Area Percentage Assessment of Cleansing Score And Computed Assessment Of Cleansing Score For Capsule Endoscopy," *Saudi Journal of Gastroenterology*, vol. 19, no. 4, pp. 160–164, 2013.

[149] J. Ju *et al.*, "Clean Mucosal Area Detection of Gastroenterologists Versus Artificial Intelligence in Small Bowel Capsule Endoscopy," *Medicine*, vol. 102, no. 6, 2023.

[150] P. Handa, N. Goel, and S. Indu, "Automatic Intestinal Content Classification Using Transfer Learning Architectures," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, 2022, pp. 1–5.

[151] J. H. Nam *et al.*, "Development and Verification of A Deep Learning Algorithm To Evaluate Small-Bowel Preparation Quality," *Diagnostics*, vol. 11, no. 6, p. 1127, 2021.

[152] J. H. Nam *et al.*, "Development of a deep learning-based software for calculating cleansing score in small bowel capsule endoscopy," *Sci Rep*, vol. 11, no. 1, p. 4417, 2021.

[153] R. Leenhardt *et al.*, "A Neural Network-Based Algorithm for Assessing The Cleanliness Of Small Bowel During Capsule Endoscopy," *Endoscopy*, vol. 53, no. 09, pp. 932–936, 2021.

[154] A. Vats *et al.*, "Learning More for Free-A Multi-Task Learning Approach for Improved Pathology Classification in Capsule Endoscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 3–13.

[155] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and

prospects," *Science (1979)*, vol. 349, no. 6245, pp. 255–260, 2015.

[156] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[157] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[158] V. S. Kodogiannis *et al.*, "The usage of soft-computing methodologies in interpreting capsule endoscopy," *Eng Appl Artif Intell*, vol. 20, no. 4, pp. 539–553, 2007.

[159] M. D. *et al.*, "Beyond Lesion Detection: Towards Semantic Interpretation of Endoscopy Videos," in *Engineering Applications of Neural Networks*, Springer. Accessed: May 27, 2022. [Online]. Available: https://books.google.co.in/books

[160] V. Rawat, A. Jain, and V. Shrimali, "Automated Techniques for The Interpretation Of Fetal Abnormalities: A Review," *Appl Bionics Biomech,* vol. 2018.

LIST OF PUBLICATIONS

Journal Papers

- P. Handa *et al.*, "Automatic Detection of Colorectal Polyps with Mixed Convolutions and Its Occlusion Testing", *Neural Comput Appl*, vol. 35, no. 26, pp. 19409–19426, 2023. (accepted and published)
- P. Handa *et al.*, "Comprehensive Evaluation of a New Automatic Scoring System for Cleanliness Assessment in Video Capsule Endoscopy", *Int J Imaging Syst Technol*, vol. 34, no. 3, p. e23097, 2024. (accepted and published)
- P. Handa *et al.*, "A Multi-Label Dataset and Its Evaluation for Automated Scoring System for Cleanliness Assessment In Video Capsule Endoscopy", *Phys Eng Sci Med*, pp. 1–14, 2024. (accepted and published)
- P. Handa *et al.*, "AI-KODA Score Application for Cleanliness Assessment in Video Capsule Endoscopy Frames", *Minimally Invasive Therapy & Allied Technologies.* (accepted and in-production)

Conference Papers

- P. Handa, N. Goel, and S. Indu, "Datasets of Wireless Capsule Endoscopy For AI-Enabled Techniques", in *International Conference on Computer Vision and Image Processing*, Springer, 2021, pp. 439–446 (accepted and published)
- P. Handa, N. Goel, and S. Indu, "Automatic Intestinal Content Classification Using Transfer Learning Architectures", in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, 2022, pp. 1–5 (accepted and published)

Book Chapters

- P. Handa *et al.*, "Computer-aided polyp detection using customized Convolutional Neural Network architecture," *in Intelligent Data Analytics for Bioinformatics and Biomedical Systems*, Wiley. (accepted and in-production)

Released Datasets

- P. Handa *et al.,* “Test data of published work titled 'Automatic Detection of Colorectal Polyps with Mixed Convolutions and its Occlusion Testing'”, Neural Computing and Applications, vol. 35. Zenodo, pp. 19409–19426, Jun. 27, 2023. doi: 10.1007/s00521-023-08762-z. (open-sourced)
- P. Handa *et al*., “AI-KODA Dataset: An AI-Image Dataset for Automatic Assessment of Cleanliness in Video Capsule Endoscopy as per KOrea-CanaDA Scores”. figshare. Dataset. May 13, 2024. doi:10.6084/m9.figshare.25807915.v1. (open-sourced)

Extended Work Done

Journal Papers

- P. Handa, H. Mangotra and N. Goel, “Effect of Selection Bias On Automatic Colonoscopy Polyp Detection”, *Biomed Signal Process Control*, vol. 85, p. 104915, 2023. (accepted and published)

Patent

- P. Handa *et al***.** (2024). A system and method to score capsule endoscopy frames using KODA [KOrea-canaDA] scoring method (Patent no. 511453). Government of India. (granted)

Biomedical Challenges

- N. Goel, B.Subramanian, P. Handa, and Deepak Gunjan. (2023). Auto-WCEBleedGen Challenge Version V1. *In collaboration* with the 8th International Conference on Computer Vision and Image Processing (CVIP) 2023, IIT Jammu. (completed)
- N. Goel, P. Handa, and D. Gunjan. (2024). Auto-WCEBleedGen Challenge Version V2. *In collaboration* with the 31st IEEE International Conference on Image Processing (ICIP) 2024, Abu Dhabi, UAE. (completed)

Conference Papers

- H. Mangotra, P. Handa, and N. Goel, “Open-Source Datasets for Colonoscopy Polyps and Its AI-Enabled Techniques”, in *International Conference on Communication and Intelligent Systems*, Springer, 2022, pp. 63–76. (accepted and published)
- T. Singh, P. Handa, and N. Goel, “Automatic GI Bleeding Detection: A Comparative Analysis of Pre-Trained Deep Learning Architectures”, in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, 2023, pp. 260–265. (accepted and published)
- T. Singh, P. Handa, and N. Goel, “Automated BBPS Scoring in Colonoscopy: A Comparative Analysis of Pre-trained Deep Learning Architectures”, in *International Conference on Computer Vision and Image Processing*, Springer, 2023, pp. 25–36.
- P. Handa, R. A. Sachdeva, and N. Goel, “CNN Architecture-Based Image Retrieval of Colonoscopy Polyp Frames”, in *International Conference on Data Analytics and Computing*, Springer, 2022, pp. 15–23.
- A. Goyal *et al.*, “Automatic Detection of WCE Bleeding Frames Using Hybrid Features And Machine Learning Algorithms”, in *2022 IEEE India Council International Subsections Conference (INDISCON)*, IEEE, 2022, pp. 1–7.
- P. Handa, R. A. Sachdeva, and N. Goel, “CNN Architecture-Based Image Retrieval of Colonoscopy Polyp Frames Check for updates”, in *Proceedings on International Conference on Data Analytics and Computing: ICDAC 2022*, Springer Nature, 2023, p. 15.

Abstract/ Poster Presentation

- P. Handa *et al.*, “Cleanliness Assessment of small bowel capsule endoscopy using AI-KODA score application*”, Korean Society of Gastrointestinal Endoscopy (KSGE) Days 2023*, September 8-9, 2023, Seoul, South Korea. (accepted and published in KSGE proceedings)

Released Datasets

- P. Handa *et al.,* “AutoWCEBleedGen-Test Dataset (Improved)”. Zenodo, Feb. 11, 2024. doi: 10.5281/zenodo.10642779.
- P. Handa *et al.*, “WCEbleedGen: A wireless capsule endoscopy dataset containing bleeding and non-bleeding frames”. Zenodo, Nov. 19, 2023. doi: 10.5281/zenodo.10156571.
- P. Handa *et al.*, “Meta-data of published work titled 'Effect of selection bias on Automatic Colonoscopy Polyp Detection'”, Biomedical Signal Processing and Control, vol. 85. Zenodo, p. 104915, Apr. 14, 2023. doi: 10.1016/j.b.
- P. Handa *et al*, “VCE-AnomalyNet: A New Dataset Fueling AI Precision in Anomaly Detection for Video Capsule Endoscopy”, Zenodo, April 2, 2024. Doi: 10.5281/zenodo.10909126.

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

PLAGIARISM VERIFICATION

Title of the Thesis: **Medical Image Analysis of Wireless Capsule Endoscopy Data**

Total Pages: **154**

Name of the Scholar: **Palak Handa**

Supervisor: **Prof. S. Indu**

Co-Supervisor: **Prof. Nidhi Goel**

Department: **Electronics and Communication Engineering**

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: **Turnitin**

Submission ID: **2249843911**

Similarity Index: **47%**

Self-Publication(s) Similarity Index: **38%**

Final Total Similarity Index: **9%**

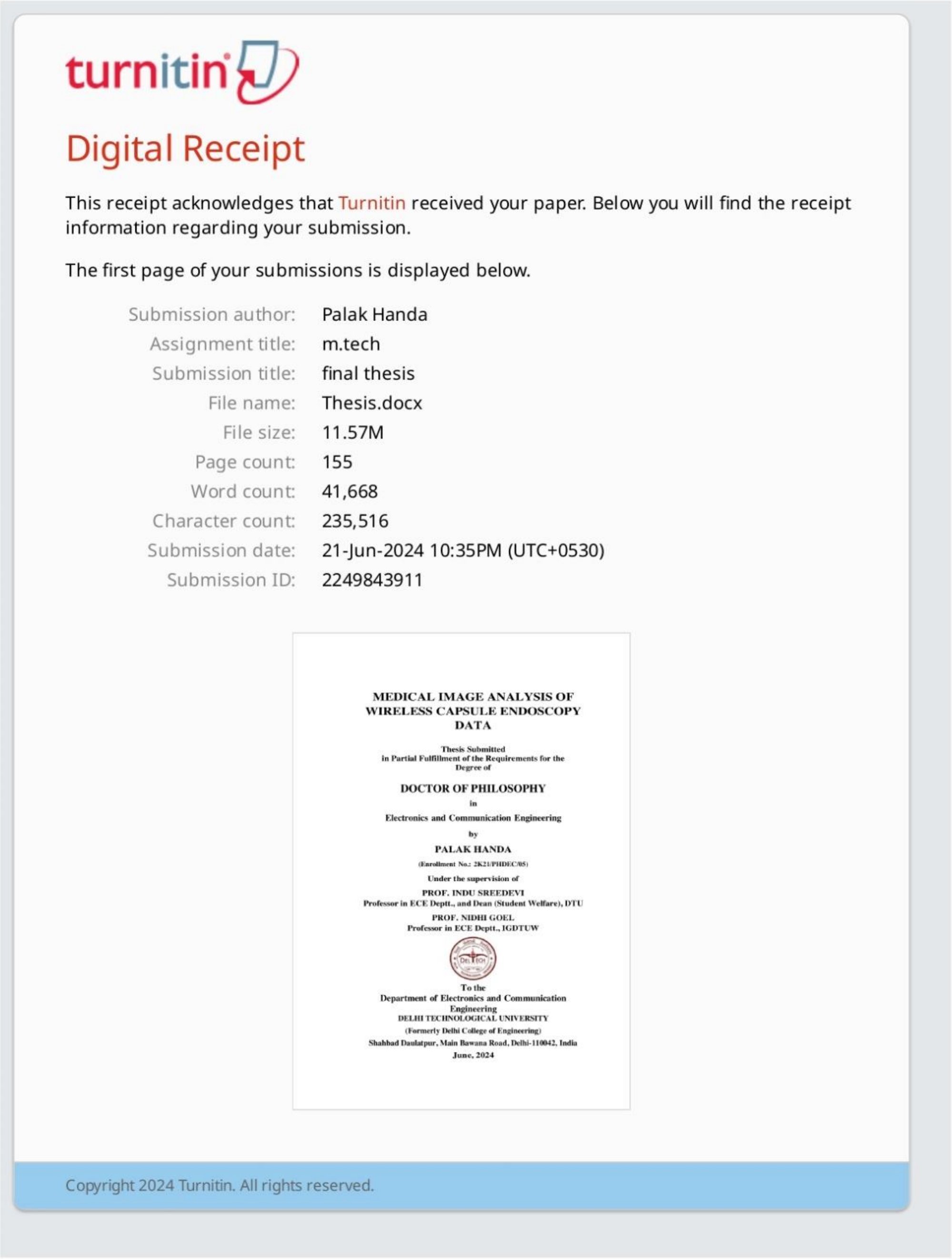
Total Word Count: **41,668**

Date: **June 21, 2024**

PalakHanda

Candidate's Signature **Signature of Supervisor(s)**

Plagiarism Report (Digital Receipt)



turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Palak Handa
Assignment title:	m.tech
Submission title:	final thesis
File name:	Thesis.docx
File size:	11.57M
Page count:	155
Word count:	41,668
Character count:	235,516
Submission date:	21-Jun-2024 10:35PM (UTC+0530)
Submission ID:	2249843911

MEDICAL IMAGE ANALYSIS OF
WIRELESS CAPSULE ENDOSCOPY
DATA

Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

in

Electronics and Communication Engineering

by

PALAK HANDA

(Enrollment No.: 2K21/PHDEC/05)

Under the supervision of

PROF. INDU SREEDEVI
Professor in ECE Deptt., and Dean (Student Welfare), DTU

PROF. NIDHI GOEL
Professor in ECE Deptt., IGDTUW

To the
Department of Electronics and Communication
Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
June, 2024

Copyright 2024 Turnitin. All rights reserved.

Curriculum Vitae/ Brief Profile of Ms. Palak Handa

Education

- **Ph.D., ECE (Computer Vision and Medical Image Processing)** 08/2021-07/2024

  Thesis title: Medical Image Analysis of Wireless Capsule Endoscopy Data
  Supervisor: Prof. S. Indu
  Co-Supervisor: Prof. Nidhi Goel
  Medical Expert: Dr. Deepak Gunjan
  Dept. of ECE, DTU, Bawana Road, Shahbad Daulatpur Village, Rohini, Delhi 110042, India
- **M.Tech., ECE (VLSI Design)** 08/2019-06/2021

  Thesis title: Age-related differences in brain metabolite (GSH and GABA) and neuro-psychological profiling in DLPFC
  Supervisor: Prof. Nidhi Goel
  Co-Supervisor: Prof. Pravat. K. Mandal
  Dept. of ECE, Indira Gandhi Delhi Technical University for Women (IGDTUW), Kashmere Gate, New Delhi 110006, India
- **B.Tech., Biomedical Engineering (Allied ECE)** 08/2015-06/2019

  Thesis title: Design of low-cost pulse Plethysmography device for health monitoring
  Supervisor: Dr. Sandeep Jaiswal
  Dept. of ECE, Mody University, Sikar Rd, Laxmangarh, Rajasthan, India

Skills

- Speaking languages: English and Hindi
- Coding: Python, MATLAB, LaTeX
- Research Interests: Medical Image and Signal Analysis, Computer Vision, Machine learning, CAD and instrumentation
- Miscellaneous: Academic research, teaching, mentoring and reviewing

Professional Experiences

- Scientist (permanent) 06/2024-present

  Research Center for Medical Image Analysis and Artificial Intelligence, Dept. of Medicine, Danube Private University, Austria
- Full-time DTU Fellow (teaching assistant) 08/2021-05/2024

  Computer Vision Lab, Dept. of ECE, DTU
- Dissertation trainee 01/2021-06/2021

  Neuroimaging and Neuro-spectroscopy (NINS) Lab, National Brain Research Center, Manesar, India
- Summer Research Intern 05/2018-07/2018

  OPTIMAL Lab, Dept. of Mathematics, IIT Indore, India

- Virtual Marketing Education Intern 12/2016-02/2017

  Pine Biotech, Texas, USA
- Biomedical Intern 05/2016-07/2016

  Dept. of Biomedical Engineering, Max Hospital Shalimar Bagh, Delhi

Research paper publications

S. No.	Entity	Accepted and published	Communicated
1.	International Journal (SCI/ SCIE)	12	1
2.	International Journal (Scopus)	1	NIL
3.	Book Chapters (Scopus Indexed)	6	1
4.	National/International Conferences	19	2
5.	Patent	2 (granted)	5 (published)
6.	Datasets	11 (open-sourced)	NIL

Awards/ Honors

- Commendable Research and Patent Award 07/2024

  Published 5 SCI-indexed journal papers and 1 granted patent in Dept. of ECE, IGDTUW Delhi. Won a cash prize.
- RIKEN CBS Summer Program 2024 Participant 07/2024

  Selected as a participant for the lecture series at Tokyo, Japan in July 2024.
- Young Investigator Award and Academic Grant Winner 09/2023

  E-poster presentation at Korean Society of Gastrointestinal Endoscopy (KSGE) Days 2023, Seoul, South Korea. Won a cash prize of $300. Financially sponsored by DTU.
- Virtual WomenTech Global Conference 2023 Student Scholarship 05/2023

  Received an alley ticket to attend the premier conference.
- Commendable Research Award 03/2023

  Published an SCI-indexed journal paper in Dept. of ECE, IGDTUW Delhi. Won a cash prize of Rs. 25,000.
- Outstanding Performer as a Mentor in Education Mentoring Program (EMP) 11/2022

  EMP is a Delhi govt. initiative to promote STEM education amongst 10-12th class girl students. Mentored 5 students for 6 months.
- Vice-Chancellor Gold Medal Recipient 10/2022

  University M.Tech. Topper in IGDTUW, Delhi.
- Virtual Grace Hopper Celebration 2021 Student Scholarship 07/2021

  One of the 250 Women in Technology to receive this scholarship across the globe.
- Best Alumni Research Paper Award 02/2020

  Conference paper presentation at ICONC3 2020, Mody University, Rajasthan. Won a cash prize of Rs. 3000.

Service

Self-Initiatives

- Brainchild (team lead), MISAHUB 08/2022-present

  Virtual platform to showcase research work, mentor undergraduate students, and promote research in medical imaging and signal analysis. Mentored +100 undergraduate and graduate students in research writing and projects since 2019.
- Research Project Mentor 09/2019-04/2024

  Computer Vision and Image Processing Lab (CVIP), Dept. of ECE, IGDTUW
- B.Tech. and M.Tech. Class Representative 08/2015-06-2021

  Administrative work and management

Organizing member

- Auto-WCEBleedGen Challenge version 2

  Organized in collaboration with 31st IEEE International Conference on Image Processing (ICIP), in Abu Dhabi from 29-31 Oct, 2024.
- Auto-PCOS Classification Challenge

  Organized in collaboration with IEEE Women in Engineering. IGDTUW, Delhi Section..+600 registrations received.
- Auto-WCEBleedGen Challenge version 1

  Organized in collaboration with 8th International Computer Vision and Image Processing (CVIP) 2023, IIT Jammu. +1200 registrations received.
- Two-week refresher course

  Topic 'Introduction to Signal, Image Processing and Machine Learning'. Organized in collaboration with IEEE DTU and Dept. of ECE, DTU Delhi for B.Tech. 2nd year students of all branches in DTU.
- ATAL FDP on 'AI and Data analytics for healthcare'

  by Dept. of ECE, IGDTUW Delhi
- ATAL FDP on 'Data Science and its Applications' *by Dept. of ECE, IGDTUW Delhi*

Peer-reviewing

- (Served as) Referee for:
  - Journals
    - IEEE: IEEE Access
    - Springer: SN Computer Science, Multimedia Tools and Applications, Scientific Reports
    - Elsevier: Biomedical Signal Processing and Control
    - Taylor and Francis: Health, Medicine and Therapeutics, Computer Methods in Biomechanics and Biomedical Engineering
  - Conferences
    - CVIP 2024, IIITDM, CVIP 2023, IIT Jammu, CVIP 2022, VNIT Nagpur
    - IEEE ICMLA 2023 (PC member), Florida, USA
    - ICONIP 2022 (PC member), New Delhi, IEEE SOLI 2022, New Delhi
    - IEEE ICMLA 2022 (PC member), Bahamas
    - BIBE 2022, China, IWEG 2022, China, ICBDS 2022, China, ICDAC 2022, China, ISCS 2022, China
    - IEEE MysuruCon 2022, Mysore
    - ICDSIS 2022, Hassan

Professional memberships

- WomenTech Global 2023-present
- International Association of Engineers (IAENG) (316406) 2022-present
- Indian Unit for Pattern Recognition and Artificial Intelligence (L-257) 2022-present
- Grace Hopper Celebration 2021-present
- IEEE Graduate Student Member 2020-2024
  (Signal Processing Society, Young Professionals, Women in Engineering) (96155761)
- Force Biomedical 2016-present

Full-time DTU Fellow responsibilities

- Teaching
  - Theory
    - Digital Electronics, summer semester, 2022
  - Lab
    - Digital Electronics, summer and fall semester, 2023-24
    - Digital Signal Processing, summer semester, 2023
    - Computer Architecture, fall semester, 2023
    - Digital Design and Verification summer semester, 2023
- Administrative work and management, examination duties

Certificates

- 6th and 7th Summer School on AI
  Center for Visual Information Technology (CVIT), IIIT Hyderabad
- Health Research Fundamentals
  SWAYAM MHRD

PalakHanda

Thank you very much for reading.

“Auron se kya? Khud hi se pooch lenge raahein.
Yaheen kaheen, mauzon mein hi, dhoondh lenge hum.
Boondon se hi to hai waheen, bandh lenge leharein,
pairon tale jo bhi mile. Baandh lenge hum kinaare, kinaare.”

Que Sera, Sera!

~ to new beginnings
Palak Handa, Ph.D.

Palak Handa   Indu Ashok