Future Generation Computer Systems 143 (2023) 191–214

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Machine learning based small bowel video capsule endoscopy analysis: Challenges and opportunities

Haroon Wahab [a], Irfan Mehmood [a,*], Hassan Ugail [a], Arun Kumar Sangaiah [b,c], Khan Muhammad [d,*]

[a] Centre for Visual Computing, University of Bradford, Bradford BD7 1DP, UK
[b] International Graduate Institute of Artificial Intelligence, National Yunlin University of Science and Technology, Douliou, 64002, Taiwan
[c] Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon
[d] Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea

ARTICLE INFO

Article history:
Received 1 August 2022
Received in revised form 7 November 2022
Accepted 14 January 2023
Available online 18 January 2023

Keywords:
Machine learning
Visual computing
Deep learning
Capsule endoscopy
Small bowel
Precision diagnostic

ABSTRACT

Video capsule endoscopy (VCE) is a revolutionary technology for the early diagnosis of gastric disorders. However, owing to the high redundancy and subtle manifestation of anomalies among thousands of frames, the manual construal of VCE videos requires considerable patience, focus, and time. The automatic analysis of these videos using computational methods is a challenge as the capsule is untamed in motion and captures frames inaptly. Several machine learning (ML) methods, including recent deep convolutional neural networks approaches, have been adopted after evaluating their potential of improving the VCE analysis. However, the clinical impact of these methods is yet to be investigated. This survey aimed to highlight the gaps between existing ML-based research methodologies and clinically significant rules recently established by gastroenterologists based on VCE. A framework for interpreting raw frames into contextually relevant frame-level findings and subsequently merging these findings with meta-data to obtain a disease-level diagnosis was formulated. Frame-level findings can be more intelligible for discriminative learning when organized in a taxonomical hierarchy. The proposed taxonomical hierarchy, which is formulated based on pathological and visual similarities, may yield better classification metrics by setting inference classes at a higher level than training classes. Mapping from the frame level to the disease level was structured in the form of a graph based on clinical relevance inspired by the recent international consensus developed by domain experts. Furthermore, existing methods for VCE summarization, classification, segmentation, detection, and localization were critically evaluated and compared based on aspects deemed significant by clinicians. Numerous studies pertain to single anomaly detection instead of a pragmatic approach in a clinical setting. The challenges and opportunities associated with VCE analysis were delineated. A focus on maximizing the discriminative power of features corresponding to various subtle lesions and anomalies may help cope with the diverse and mimicking nature of different VCE frames. Large multicenter datasets must be created to cope with data sparsity, bias, and class imbalance. Explainability, reliability, traceability, and transparency are important for an ML-based diagnostics system in a VCE. Existing ethical and legal bindings narrow the scope of possibilities where ML can potentially be leveraged in healthcare. Despite these limitations, ML based video capsule endoscopy will revolutionize clinical practice, aiding clinicians in rapid and accurate diagnosis.

1. Introduction

The gastrointestinal (GI) system, commonly known as the digestive system, plays a vital role in sustaining the health of the human body to perform daily activities. A large-scale multinational study [1] suggested that over 40% of the people worldwide are affected by GI disorders. GI-related cancers account for 26% of the total cancer cases and 35% of all cancer-related deaths worldwide. Furthermore, gastrointestinal diseases can manifest at least once in an individual's lifetime in 43% of the UK population [2]. Therefore, early diagnostic check-up is essential and effective for ruling out fatal medical conditions and taking in-time therapeutic measures in case malignancy or disease is diagnosed. The GI tract comprises two parts: the upper and lower GI tract. The upper GI tract is generally considered to be the mouth, esophagus, stomach, and the first part of the small intestine (duodenum).

* Corresponding authors.
E-mail addresses: I.Mehmood4@bradford.ac.uk (I. Mehmood), khanmuhammad@g.skku.edu, khan.muhammad@ieee.org (K. Muhammad).

https://doi.org/10.1016/j.future.2023.01.011

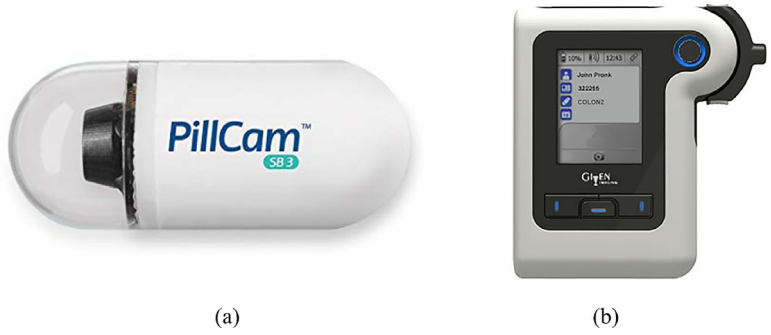

Fig. 1. An example Video Capsule Endoscopy. (a) Medtronic PillCam SB3 capsule and (b) Medtronic PillCam recording device (image courtesy of Medtronic [6]).

The lower GI tract extends from the small intestine to the large intestine. Endoscopy and colonoscopy are interventional diagnostic procedures used for the upper and lower GI, respectively. A thin flexible probe comprising a camera, light source, and therapeutic port is inserted into the body to capture a live video of the GI organs, which is displayed on the screen for real time analysis. However, this procedure requires the expert to be highly adept and vigilant during the operation; moreover, diagnostics must run in parallel. Despite the expert-based procedure, the misdetection rate for colonoscopic polyps is 20% [3]. Moreover, an invasive procedure is associated with complications such as perforation, bleeding, and infection during colonoscopy [4]. In particular, an invasive procedure, along with its risks and fear of being at the hospital, naturally deters patients in need of early diagnosis. An alternative noninvasive technique was introduced in the early 2000s [5], called video capsule endoscopy (VCE) or wireless capsule endoscopy (WCE). A miniaturized camera, light source, small electronics, and transmitter antenna are embedded in a pill that can be easily swallowed by the patient. The capsule captures the video frames as it traverses passively along the GI tract and sends frames wirelessly to a receiver device held by the patient. The typical frame capture rate varies in the range 1–30 frames per second, and 50–100 thousand frames per patient can be captured. The capsule is disposable and passes out with feces after remaining inside the body for approximately 10–16 h. The video is then transferred to a computer for analysis by an expert. Fig. 1 shows a Medtronic [6] VCE capsule and recording device. Such devices can potentially facilitate early diagnosis of certain diseases such as inflammatory bowel diseases (IBD). IBDs are highly associated with the development of malignancies over time; if not detected early, they may progress to lethal diseases such as cancer [7,8]. Therefore, the early detection of IBD is essential to prevent cancer in the digestive organs. Capsule endoscopy is a non-invasive, patient-friendly, and hospital-free procedure and is an apt candidate for delivering an early diagnostic or screening check for IBDs on a massive scale. In other words, capsule endoscopy, as an appropriate screening procedure among masses, can potentially reduce cancer cases by detecting cancer-risk conditions such as IBD early. Furthermore, VCE can potentially carry out remote screening and diagnostics in areas where experts are not readily available for more complex interventional diagnostic procedures. However, what offers patients a hospital-free, risk-free, and friendly experience burdens experts with the task of viewing these VCE videos, which comprises thousands of frames, offline. At times, a pathological condition might be present in just one frame, which could lead to an incorrect diagnosis. Typically, the time taken for an expert to analyze a VCE video is approximately 1–2 h [9].

Artificial intelligence (AI) is considered as new electricity that transforms every field of life, and hence, holds potential in coping with the analysis of substantial amounts of VCE frames. AI has advanced considerably in the last two decades, particularly in the fields of computer vision and natural language processing (NLP). Machine learning, which is the backbone of AI, encompasses computational methodologies and algorithms that enable machines to learn knowledge and patterns from either data or their own experience. Until the late 2000s, image and video analysis was mostly performed by designing handcrafted filters to extract useful features from raw images or video frames. The extracted features were subsequently fed into conventional machine learning methods such as support vector machine (SVM), random forests, and logistic regression for downstream tasks such as classification or segmentation. Deep neural networks (DNN) and deep learning (DL) can extract useful features directly from raw data and have achieved considerable success with the availability of parallel computing hardware in the last two decades. Deep convolutional neural networks (DCNNs), a type of DNN for images and computer vision in deep learning, have achieved tremendous milestones in image classification, detection, segmentation, and tracking [10]. The basic architecture of CNNs is inspired by the primate visual cortex [11,12]. The stack of learnable convolutional filters extracts features in their local receptive fields, which are further subjected to nonlinear activation functions and local pooling to cater to the abstract concepts in successive layers. Earlier layers extract low-level features such as edges, color, gradient orientation, and texture, whereas successive layers learn more high-level features such as shape, entities, and the relationship between them. Learning in CNNs is achieved by employing the widely-used back propagation as in multi-layer perceptron (MLP) and other artificial neural networks (ANN). Since the inception of the first CNN architecture, many modifications and enhancements have been proposed in different aspects, such as skipping the connections, branching within a layer, modifying a processing unit, and optimization strategies for hyperparameters. Several alternatives to CNNs have also been proposed, such as graph neural networks (GNN) [13] or capsule neural networks (CapsNet) [14] for image analysis. However, CNNs are reported to be the most successful and are widely used in real-life computer vision applications [10].

Earnest attempts have been made to leverage the true potential of AI in VCE analysis and precision diagnostics. In this study, we critically analyze and present a survey of those attempts; that is, we highlight both the current challenges VCE poses to AI as a computer vision problem in the context of machine vision and the essential requirements for a prospective AI solution in the light of recent articles and reports published by gastroenterologists. We focus on video capsule endoscopy, particularly for the small bowel, as VCE is the gold standard and first-line examination for small bowel disorders [15]. Various

interesting surveys have been conducted on AI for VCE analysis. A clinical survey presented in Kim and Lim [16] discussed recently proposed deep learning-based solutions for identifying various small bowel diseases and concluded that AI in VCE is still in its research phase, and with rapidly evolving methods, a fully automated VCE analysis may prove helpful in GI diagnostics. Trasolini et al. [17] presented a clinical review of different AI-based VCE studies conducted between 2000 and 2020. Deep learning studies of VCE are all retrospective in nature and are highly prone to bias. Studies have discussed the future need for a generalized AI-based system for VCE [18–21] . Shelly et al. [15], in their meta-analysis and systematic review, presented a review of CNN-based VCE methods. Furthermore, they also performed quality assessment and bias risk analysis on existing studies. Khan et al. [22] presented a survey of several computer-vision-based strategies and tasks employed in the VCE domain. Moreover, they highlighted challenges, recommendations, and future directions, with a major focus on the prospective potential of delivering a smart healthcare diagnostic system by merging the Internet of Things (IoT) and VCE.

Comparing the existing CNN-based detection approaches suggest higher accuracy, sensitivity, and specificity for single anomaly detection tasks, such as bleeding, angioectasia, and ulcers. However, many of these studies were retrospective in nature and had an intrinsic risk of bias. Furthermore, many studies did not describe their data dynamics and many are not publicly available. Studies on detecting multiple abnormalities are rare. Therefore, there is a dire need to apply these AI methods prospectively to a large multicenter dataset in a more pragmatic way to deliver an efficacious impact in the field of video capsule endoscopy. In this study, contemporary challenges are identified hampering AI to be effectively applied in VCE domain. VCE frames are intrinsically demanding from a computer vision perspective. In contrast to conventional endoscopic probe cameras, a capsule is untamed and wild in its motion-capturing frames in a highly diverse and concealed manner (to the examiner) [23]. This, in addition to the low resolution of VCE frames compared to endoscopic frames, requires extra care to be taken in the context of computer vision. Deep-learning approaches have outperformed conventional handcrafted feature-based methods. However, when it comes to robustness and generalizability, paramount factors for employing AI in clinical practice, deep learning is deemed to be data-hungry [24]. The availability of datasets for medical imaging has been a persistent challenge, let alone the datasets for VCE, where labeling tens to hundreds of thousand frames for a single patient is both a time- and attention-seeking task. In this context, automated AI-based labeling procedures and un-supervised or semi-supervised learning offer an opportunistic area to explore. Merging different datasets to build a large multicenter dataset offers new challenges in resolving data incongruity in the VCE domain, where the choice of defining frame labels is highly disparate. Coping with diversity, introduced by combining frames from different manufacturers, is another challenge. Rare medical conditions should not be underrepresented in data-driven learning processes. More often, a rare condition happens to be more fatal; for example, polyps are rare in the small bowel but very significant in detecting the possibility of malignancy. The competitive performance metrics reported in the lierature for selective tasks on selective data in a retrospective way suggests considerable potential; however, developing an end-to-end holistic solution to harness the essence of AI in video capsule endoscopy might first need these contemporary challenges to be addressed.

The remainder of this paper is organized as follows: Section 1 introduces the concept of AI for VCE analysis and elucidates its potential impact on early and precise diagnosis and large-scale remote screening. A brief overview of recently published survey articles in this field is provided. Furthermore, the challenges and recommendations for improving existing studies have been discussed. Section 2 develops a hierarchy of essential steps required for analyzing the raw frames of VCE for the diagnosis made at the disease level. The taxonomic organization of frame-level findings and their relative pertinence with disease-level findings are discussed in light of recent international consensus. Semantic and visual descriptions of typical representative frame-level findings are also presented in this section. Section 3 provides a brief overview and comparison of existing datasets for small-bowel video capsule endoscopy. Section 4 critically describes existing machine learning and computational methods applied to several tasks in the analysis of VCE, such as summarization, classification, segmentation, and localization. Section 5 reflects on contemporary challenges identified as hampering factors in reaching the full potential of computer-aided diagnosis for VCE. Possible opportunistic areas for coping with these challenges are also highlighted. Finally, Section 5 concludes the study.

2. Overview of small bowel diseases in context of VCE

The small bowel, a tubular structure physically located in the middle of the GI tract before the large intestine and after the stomach, is responsible for 90% of the digestion- and absorption-related workload [25]. In adults, the average length of the small bowel is approximately 6–7 m, is highly convoluted, and narrowly twisted around the abdominal cavity [26]. Longitudinally, it is further divided into three parts: duodenum, jejunum, and ileum. The wall of the small bowel is layered cross-sectionally, with the inner-most layer containing small finger-like projections extending into the lumen, called villi, which increases the surface area of the intestinal wall for maximum absorption. Narrowly tangled placement of the small intestine offers a difficult approach for the traditional endoscopic probe to maneuver in a diagnostic procedure [27]. By contrast, a capsule can easily reach and access all areas of the small bowel owing to its small size. Therefore, easy accessibility to reach the entire small bowel, non-invasiveness nature, and patient friendliness make capsule endoscopy the gold standard diagnostic procedure for the diagnosis of small bowel diseases.

In traditional endoscopy, real-time endoscopic motion-controlled viewing lasts for 20–30 min; however, the analysis of an 8–12 h VCE video from diagnostic intention is a slightly different task. For example, an expert performing endoscopy is naturally more vigilant looking for a suspected anomaly, since it is at his discretion to focus and explore anything that is thought to be an anomalous area in real time. Moreover, controlled camera motion and therapeutic add-on reinforce the active attention mechanism [28–30]. In contrast, VCE video analysis is naturally a less attention-luring task owing to the following reasons: a sense of being offline and not in real time, the idea of observing a lengthy monotonous screen, low resolution of VCE frames, and less presence in the scene due to lack of motion control [31–34]. To present this offline and wearisome video analysis in a goal-oriented task manner, Fig. 2 shows a hierarchy of the VCE-based small bowel disease diagnosis process with a brief comparison of human experts and artificial intelligence in performing the tasks. First, the video frames in the raw VCE video are labeled with contextually relevant frame-level findings. This further involves subtasks such as first removing redundant, poor quality, and duplicate frames. An expert performing this task must be aware of the well-defined consensual nomenclature of contextually relevant frame-level findings in an SB VCE scope. Excessive supervised training hours are required for such kinds of image recognition tasks; however, it is relatively more challenging to classify VCE frames accurately owing to poor quality

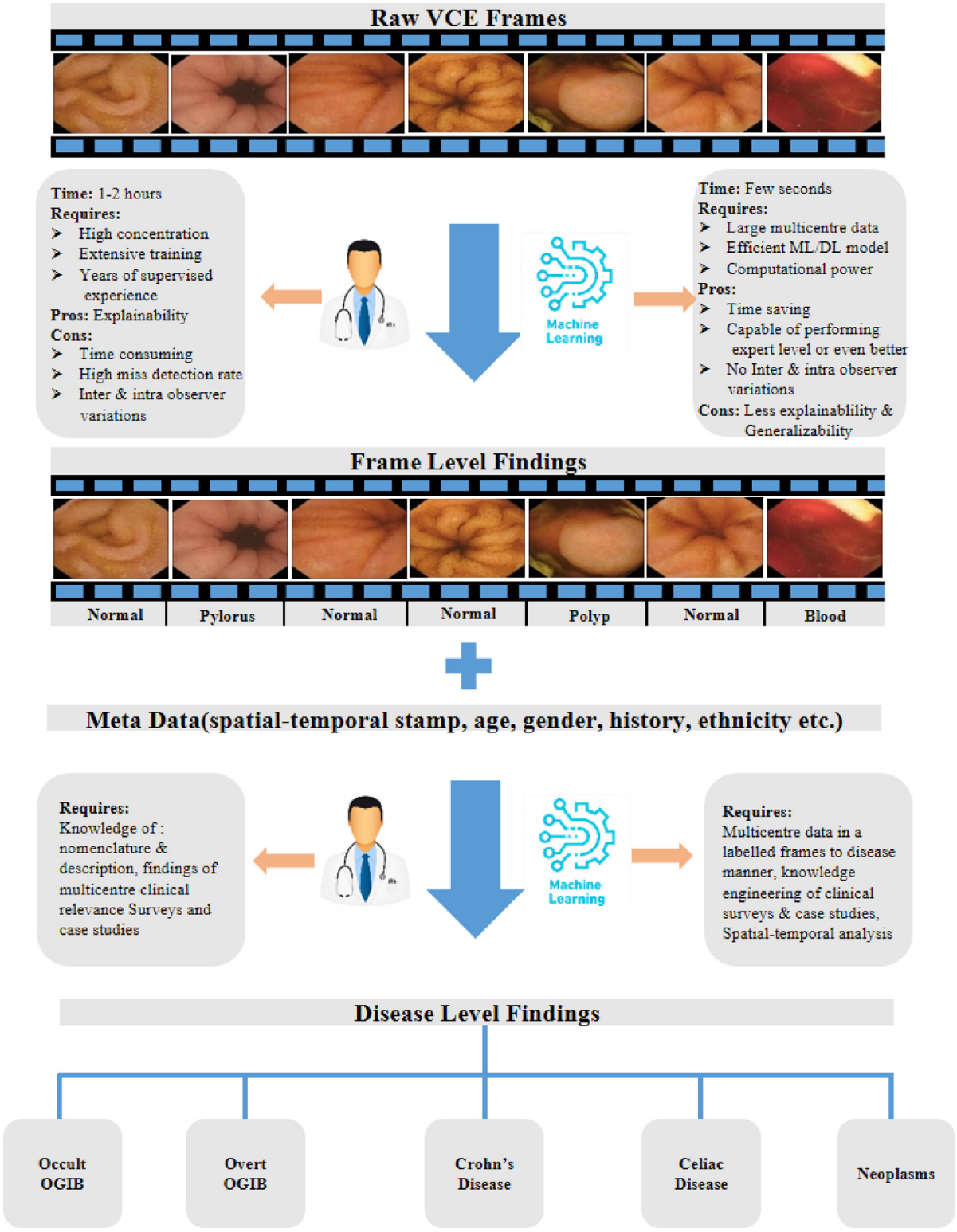

Fig. 2. Hierarchy of steps for a disease level diagnosis in small bowel Video Capsule Endoscopy.

images, subtle manifestation of anomalies, and mimicking nature of various frame-level findings. The next major and final task in this hierarchy is to develop an inference regarding disease-level finding for only those video samples comprising one or more anomalous frame-level findings. This task involves the analysis of annotated frames combined with meta-data (capsule spatial–temporal stamp, age, sex, history, ethnicity, geo-graphical location, etc.) to give a probability score for prospective small bowel digestive disease. An anomaly detected in just one frame may not be the most representative when it comes to disease-level diagnosis. In the whole video, there could be multiple occurrences of a single type of anomaly or multiple anomalies

at different places; therefore, a holistic approach catering to all possible scenarios needs to be taken in making a disease-level diagnosis. An annotated video with anomalous frames along with meta-data needs to be further analyzed and consulted for a disease-level investigation and diagnosis in light of a knowledge base, including clinical relevance studies, case studies, and domain knowledge. In practice, clinical experts mostly appear to follow the same hierarchical approach; unfortunately, AI has not been applied in such a hierarchal manner in this domain. Mostly, the task of frame-level anomaly detection has been confused with disease-level detection. However, these two tasks complement each other to reach an accurate diagnosis. For an end-to-end solution, AI needs to be applied in the same cognitive hierarchy as human experts to perform a disease level diagnosis by first analyzing each frame, and holistically analyzing all anomalous frames in a video along with other metadata to obtain a more accurate and robust disease level decision. In this section, we discuss the frame-level and disease-level findings for small bowel VCE in detail.

2.1. Frame level findings

The task of developing frame-level recognition expertise either by human experts or AI first and foremost requires a well-defined semantic and visual nomenclature for the findings under observation. Even the relatively simple case of recognizing a cat or dog first requires the visual definitions of these respective classes to be made implicitly in the visual cortex, as well as in a CNN-based model. In a more complex case, such as in VCE frame-level recognition, where the closely spaced visual peculiarities of these findings are not so conspicuously pronounced, the visual and semantic description of frame-level findings needs to be explicitly delineated and endorsed by domain experts worldwide. Since the birth of capsule endoscopy in 2000, there has been a prevalent dissensus and confusion regarding the nomenclature of frame-level findings. Korman et al. [35] proposed capsule endoscopy structured terminology (CEST) in 2005; however, it could not be adapted globally in clinical practice because of its complex structure [36]. This has been a significant setback for both VCE and AI in VCE to reach their full potential in clinical practice. The lack of consensual nomenclature led to various names and definitions of frame-level findings, specifically for ambiguous and subtle lesions [37]. Low-resolution images and diverse capturing conditions introduced by the free nature of the capsule make lesions and ulcerative conditions appear visually inseparable and less distinct than the more unequivocal frame findings for active bleeding and deep ulcers [38].

More recently, gastroenterologists and experts from different parts of the world have attempted to reach an international Delphi consensus statement on the nomenclature and semantic description of vascular, inflammatory and ulcerative, and lymphatic lesions in small bowel VCE [39,40]. This international consensus is highly commendable. Indeed, it was much needed to overcome many problems in the VCE domain. In light of this recent consensus and other related work [41–48], we propose a taxonomy of frame-level findings to help better understand the task of VCE frame annotation and small bowel-related luminal conditions for capsule endoscopy, as shown in Fig. 3.

A frame is perceived to contain significant information if an expert or AI opinion is more inclined towards one of three clusters: anatomical, pathological, and normal/normal. Normal is also a cluster and frame-level finding at the same time. Normal variants are often confused with ambiguous lesions or ulcerative conditions. Anatomical frame level findings are used as landmarks to ensure that the capsule has passed a certain region as well as in accurate localization of the capsule, and hence, are very significant to detect. The pathological cluster subsumes all types of anomalous frame findings that may collectively relate to the disease level diagnosis in the next hierarchical step. Fresh blood is a frame-level finding, however, it can coexist with various types of lesions or other anomalous conditions, resulting in different disease-level diagnosis in different individuals. Most anomalous conditions exist under the label of 'Lesions'. Indeed, lesions are a very generic term and are related to abnormal tissue changes due to injury or a certain disease. Vascular lesions are abnormal lymphatic vessels of the gastrointestinal tract that can be inherited or acquired. Inflammatory and ulcerative lesions are abnormal immune or inflammatory responses, with infiltration of inflammatory cells into the intestinal wall and manifest in various forms. Mucosal atrophy is associated with anatomical changes in the mucosal lining, which affect the absorption function of the small bowel and lead to malabsorptive diseases. Hookworms are parasites that attach to the intestinal wall and cause bleeding or small bowel infection. Any frame is non-significant if the image is blurred, lossy, poor light conditions, reduced view, or contains foreign bodies such as food content, bubbles, or feces. A brief semantic and visual description of the significant frame-level findings in Figs. 4 and 5 is presented for understanding purpose.

The pylorus is a pinkish mucosal landmark around a smooth, dark, round opening from the stomach into the first part of the small bowel. The ampulla of Vater is a landmark formed by the fusion of the pancreatic duct and the common bile duct, and resembles a trumpet-mouth-like dilated opening at the duodenal wall. The ileocecal valve is a valve formed by two-fold of the mucous membrane at the opening of the ileum into the large intestine and is marked by the presence of a group of longitudinal ileal lines forming a rosette-type pattern at the ileocecal valve. Blood is probably the easiest finding to detect bright or dark-red-colored patches over the mucosal surface. Hookworms are appeared as off-white, thread-shaped parasites attached to the mucosal wall. In mosaicism, a mosaic-like pattern can be observed on the mucosal wall. Scalloping involves a scallop pattern on the edges of mucosal folds. Small-sized granules or micronodules are formed along the mucosal surface in the granular mucosa. Minute finger-like projections called villi are reduced on the mucosal folds in the flattened mucosa. A bright-red, flat lesion with clustered and convoluted capillary dilations is a typical description of angiectasia or angiodysplasia. A small (few millimeters) flat reddish area without any manifestation of vessels surrounded by intestinal villi is a consensual description of an erythematous patch.

Red Spot/Dot is an extremely small (less than 1 mm) bright-red-colored dotty appearance, similar to a flat lesion without any vessel appearance within the mucosal layer. Phlebectasia is a bluish venous dilation that grows slightly below the mucosa. Diminutive angiectasia is the appearance of bright red non-clustered capillary dilations organized in a linear fashion with clear demarcations. Aphthoid erosion is typically a whitish center surrounded by a red nimbus with an associated loss of epithelial layering on a small scale. Deep ulceration usually manifests as a candid deep loss of tissue surrounded by a swollen mucosa with a white base. Superficial ulceration is a loss of tissue, however, to a lesser degree, between aphthoid erosion and deep ulceration. Stenosis reflects narrowing of the intestinal lumen circumference and is usually associated with a delayed passage of the capsule. Edema is defined as the appearance of enlarged, swollen, or congested villi. Hyperemia is a condition in which the villi are overly reddish. Denudation is defined as a reddish mucosal area without villi. Lymphangiectasia is diffusively elongated and the circumferential mucosa is covered with whitish enlarged villi. Chylous cysts are yellow, soft submucosal lesions that are diffused in appearance with an occasional presence of vascular

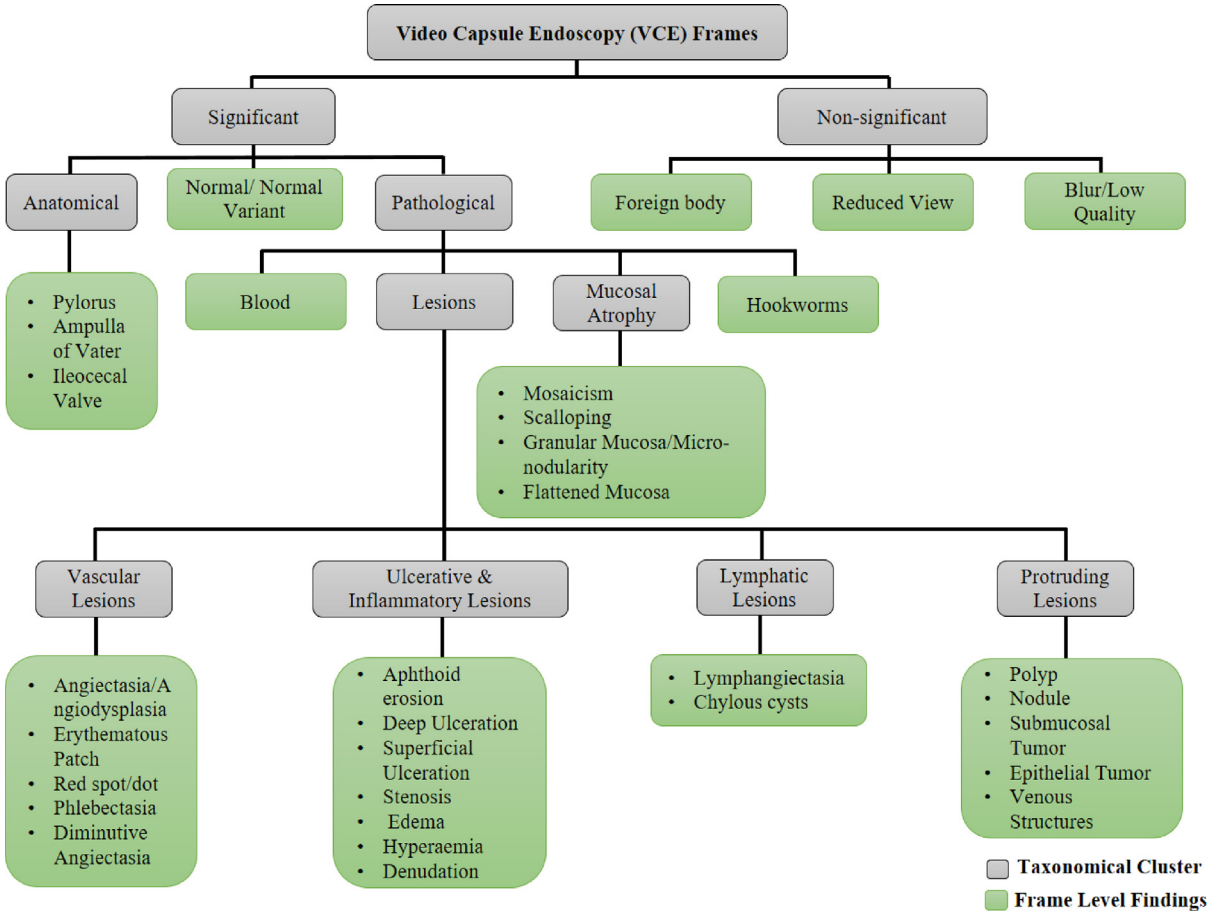

Fig. 3. A taxonomy of frame level findings for small bowel VCE frames.

patterns across the surface. Polyps and tumors manifest as mass-like protruding tissue growths, often bulging as mushroom stalks with variations in size. Nodules are typically described as white-scattered polypoids spread across patches on the mucosal surface.

The proposed taxonomy of frame-level findings may potentially leverage multiclass classification tasks [50]. A SoftMax function applied at the frame-level finding level in the hierarchy during training generates probabilities against each finding. At inference time, these probabilities can be summed all the way up in the hierarchy to give the probability for each taxonomical cluster. Therefore, pathological and visual similarities within a cluster can be contrived by considering the probabilities for cluster-level classes. This approach may potentially achieve partial summarization in VCE as a by-product of classification, considering probability scores at the highest cluster level "Significant/Non-significant" would essentially dictate the decision of informativeness about a certain frame.

2.2. Disease level finding

Annotated frames with significant frame-level findings need to be canvassed in a spatial–temporal format with estimations of the spatial position of each frame within the small bowel as well as an estimate of the relative time elapsed at a certain position. These estimated spatial–temporal stamps for significant frames are manifested as 'Meta Data' in Fig. 2, along with other patient-related metadata. An expert gastroenterologist or a prospective AI solution would assess for a disease-level diagnosis by meticulous perusal of all the above-mentioned contents. Unfortunately, the significance of this second stride in the hierarchy presented in Fig. 2 has been overlooked by researchers applying AI to VCE analysis. However, a literature review of clinical surveys and relevance analysis strongly suggest this approach. For example, stenosis narrows the lumen in its inner circumference; however, it is recommended to reinforce this finding by correlating with associated probable delays in capsule passage through this area [51]. Similarly, a certain finding in two different areas of the small bowel may lead to a different clinical relevance. Lesions in the small bowel are not related to small bowel VCE-indicative diseases in a single fashion. Most clinically relevant lesions are reported to be associated with multiple diseases. This meta-data enables physicians to contemplate coherently with their knowledge to obtain an outcome of correct diagnosis. Fig. 6 shows a clinical relevance relationship between several frame-level findings and disease-level findings, considering an international consensual clinical survey. The most relevant findings are those with single or multiple occurrences that are highly thought to be correlated with a particular disease in an international consensus [41]. Furthermore, moderately and mildly relevant are findings having moderate and mild levels of occurrence against a certain disease, respectively. Next, we briefly describe each disease level finding.

Crohn's disease is an inflammatory bowel disease (IBD) causes inflammation and lesions in the small bowel. The most relevant frame-level findings were deep ulcerations and stenosis. Moderately relevant findings included aphthoid erosion and superficial ulceration. Occult obscure gastrointestinal bleeding (OGIB) is a type of GI bleeding caused by lesions in the small bowel and

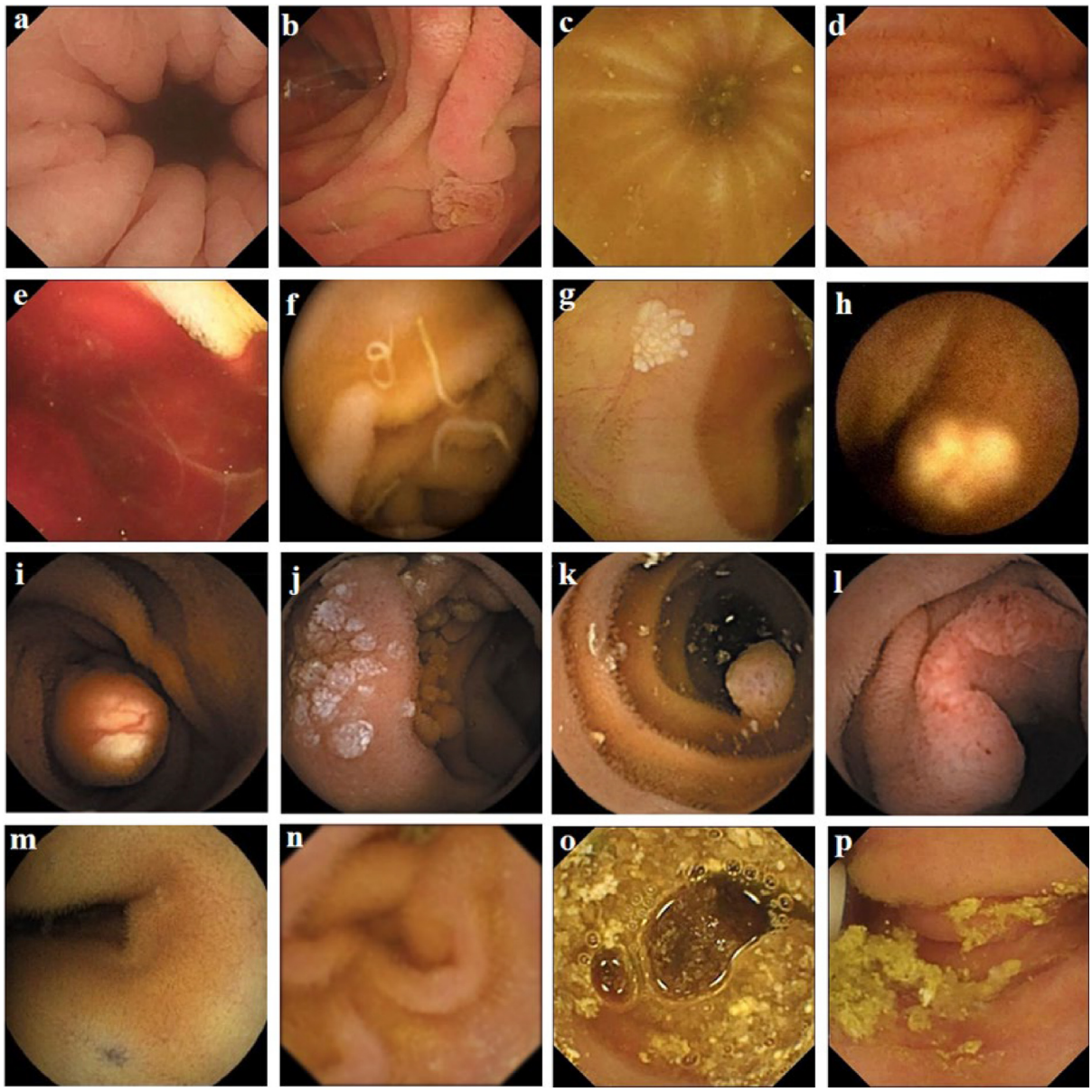

Fig. 4. Representative images of Frame Level Findings (Part-1): (a) Pylorus, (b) Ampulla of Vater, (c) Ileocecal Valve, (d) Normal, (e) Blood, (f) Hookworms, (g) Lymphangiectasia, (h) Chylous Cysts, (i) Submucosal Tumor, (j) Nodule, (k) Polyp, (l) Epithelial Tumor, (m) Venous Structures, (n) Blurred, (o) Reduced View, (p) Foreign Body.
Source: Images taken from [23,49].

occurs with or without a positive fecal occult blood test. It is often referred to as iron deficiency anemia (IDA). The most pertinent frames for occult OGIB are the blood, deep ulceration, typical angiectasia, and stenosis.

Obscure gastrointestinal bleeding (overt OGIB) is a type of GI bleeding in the small bowel that is clinically perceivable and often recurs or remains in patients despite negative initial endoscopic evaluations. The most pertinent frames were the same as those in the case of occult OGIB. Celiac disease is associated with mucosal atrophy, a condition where the immune system attacks the mucosal surface tissues of the small intestine, resulting in nutrient absorption. Although, to the best of our knowledge, no large Delphi consensus statement exists regarding the clinical relevance of this type of disease; however, a small-scale consensus [43–45,47] suggests that the following frame findings are most relevant: mosaicism, scalloping, flattened mucosa, and granular mucosa. Neoplasms are abnormal growths of cells or tissues in the small intestine that may or may not be malignant. Relevant findings include different types of protruding lesions, such as polyps, nodules, and tumors.

3. Existing VCE datasets

Since the birth of the VCE, several datasets for computer-aided analysis and training purposes have been imparted to the literature. A brief comparison of these datasets is presented in Table 1. Dataset availability and access in the medical domain is generally a prevalent problem that faces many ethical and privacy issues, let alone VCE, which has additional sparsity and redundancy challenges, thereby fostering a phenomenal requirement of much larger raw datasets for achieving effective and robust generalization. At present, most of these are not publicly available for open

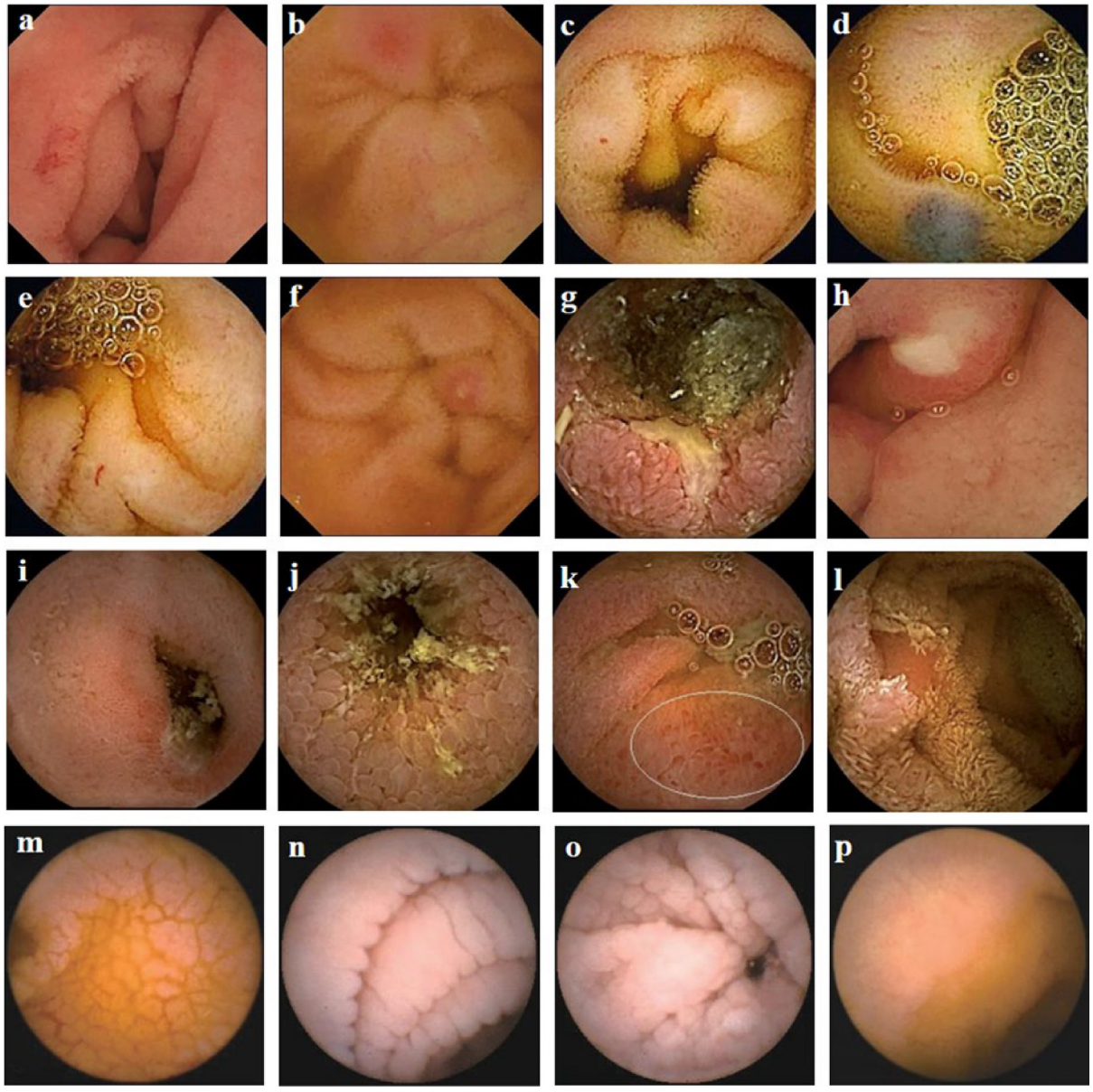

Fig. 5. Representative images of Frame Level Findings (Part-2): (a) Angiectasia, (b) Erythematous Patch, (c) Red Dot/Spot, (d) Phlebectasia, (e) Diminutive Angiectasia, (f) Aphthoid Erosion, (g) Deep Ulceration, (h) Superficial Ulceration, (i) Stenosis, (j) Edema, (k) hyperemia, (l) Denudation, (m) Mosaicism, (n) Scalloping, (o) Granular Mucosa, (p) Flattened Mucosa.
Source: Images taken from [23,49].

academic access. One commendable contribution was recently attributed to the literature (Kvasir-Capsule) [49] for their easily accessible large dataset, and comprised 117 videos and 4,694,266 images. Although only a small portion is labeled into 14 different frame-level findings, the unlabeled portion can be used for unsupervised learning or labeling by a third-party group. The label choices for different frame-level findings are non-overlapping across different datasets, posing an additional overhead in merging various datasets. Only the KID [52] dataset was labeled according to a consensus terminology standard (CEST;2005 [35]). These existing datasets and future contributions need to comply with the most recent terminology consensus [40] developed by the international community to enable the merging of datasets across the globe into a large dataset as required by a prospective DL solution. The paucity of datasets for capsule endoscopy potentially invites the exploitation of generative adversarial networks (GANs) and other synthetic data generation methods and is further explored in Section 5.2.

4. Existing computational methodologies for analysis of VCE

Analyzing capsule endoscopy videos is a crucial task for both experts and researchers who are developing intelligent analysis algorithms. Since the birth of this revolutionary diagnostic technology, a plethora of computational methods and techniques have been employed to simplify the process of analyzing a VCE video into an easy job. The focus has been on two key factors: (a) reducing reading time and (b) increasing readability by assisting in the recognition and indication of anomalies. However, in the context of computer vision, the analysis of VCE frames can be further

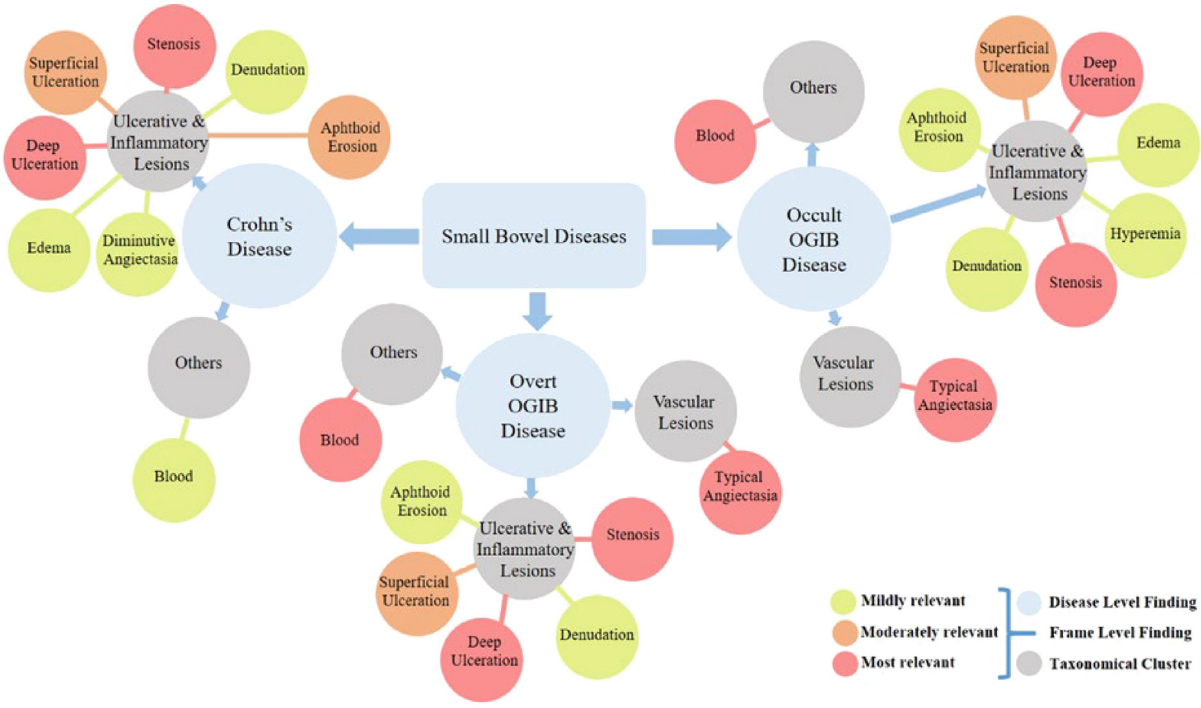

Fig. 6. Pertinence between Frame Level Findings and most indicated small bowel diseases for VCE.

Table 1
A brief overview of existing small bowel Video Capsule Endoscopy datasets.

Dataset	Frame level findings standard	Size	Ground truth	Availability
KID [52]	CEST based labeling	2371 images, 47 videos	Labels, graphical annotations	On request
GIANA [53]	Small bowel lesions-based labeling	8262 images, 38 videos	Labels	On request
CAD-CAP [54]	Labels: vascular, blood, ulcero-inflammatory, normal	25,000 images	Labels, graphical annotations	On request
Gastrolab [55]	Crohns disease and various lesions	Few hundred images	Labels	Publicly available
Kvasir-Capsule [49]	14 different frame level findings	4,694,266 images, 117 videos	Labels and graphical annotations for 47,238 images	Publicly available

categorized into three tasks: summarization, recognition, and localization. The first two tasks lie purely in the computer vision domain; however, localization of the capsule inside the body also considers other modalities, such as radio frequency (RF), motion sensors, and magnetic sensors. Before the recent burgeoning in DL, most analytic enhancements were handcrafted features based on image and signal processing techniques. However, in the last decade, much work has been done by employing deep-learning-based learnable feature methods. In this section, we present a succinct and critical survey of notable existing research methodologies and discuss state-of-the-art software enhancements that are contemporary in the clinical practice of VCE.

4.1. Summarization

Video summarization is a challenging task in multimedia systems and computer vision. In the VCE context, it becomes more challenging owing to the capsule's inapt capturing profile and associated likeliness of subtle manifestations of the anomaly. The apparent trade-off between the miss rate and extent of summarization is a unique challenge in capsule endoscopy and other anomalous behavior detection videos. Regardless of how efficiently and robustly AI-based diagnosis is generated by a prospective clinical solution; a gastroenterologist would always like to view the video himself to ensure nothing is overlooked by a computer-aided diagnosis (CAD) system. Therefore, summarization, or in a less strict sense, removal of redundant frames, becomes a rudimentary requirement for saving the reading time of an expert. The effect or extent of summarization should be optimal to avoid missing any information conducive to an otherwise anomalous condition. Therefore, the performance of summarization approaches in VCE is generally measured in terms of reading reduction time or frame reduction percentage, along with the trade-off of the associated misdetection rate.

Soon after the approval for VCE use in clinical practice, a software-aided redundancy removal need was felt by gastroenterologists and manufacturers. The Given Imaging Inc. (Medtronic now) launched initial versions of software enhancement such as Quick View (also called Rapid mode) for summarization and Suspected Blood Indicator (SBI) for indicating blood-related anomalies in the early 2000s [56]. However, prospective clinical testing has reported a significant misdetection rate as a by-product of a

reduction in reading times [57]. Similarly, Olympus introduced a summarization enhancement in the last decade condensing the entire video into approximately 2000 key frames; however, the method only detects explicit visual changes among images to decide for key frames, and anomalies belonging to small-scale visual patterns were at risk of missed detection [58]. A brief comparison of attempts made by researchers for VCE summarization is presented in Table 2.

Redundancy in CE is caused by the uncontrolled nature of the capsule, irregular motility profile under the influence of peristalsis, poor lighting conditions, reflections and blurredness, and obscurity caused by bubbles, chymes, and foreign bodies. Earlier attempts were mostly handcrafted feature-based techniques, where color space, texture-based features, and shape-based features were used. The summarization method generally involves segmentation of a video into several shots and subsequently selecting key frames from such segments based on certain criteria on the extracted features. Iakovidis et al. [59] proposed an unsupervised technique based on matrix factorization for clustering and orthogonality criteria to select key frames from a cluster. However, this way may result in grouping of timely irrelevant frames to be grouped together, eventually causing an improper selection pool for key frames. Researchers in Liu et al. [58],Lee et al. [60] and Lee et al. [61] attacked redundancy removal by estimating capsule motion based on changes in consecutive scenes and ego-motion analysis of capsule caused by peristaltic forward, backward motion of capsule respectively. However, the peristalsis cycle itself varies among different individuals, and even within an individual. Hence, the estimation of peristaltic motility using optical flow methods at a low frame rate, such as in VCE, 2 fps, may cause much visual deformation over the extent of optical flow techniques to correctly estimate the motion flow [62]. In our earlier work, we used color space transform, multi-scale contrast, moments, curvature, JD divergence, and Boolean series correlation to create saliency maps and key frame selection [63–65]. Deep learning has also been applied by several researchers. Mostly, a hybrid approach is adopted, comprising DL-based feature learning and conventional ML based methods for shot segmentation and key-frame extraction. CNNs and autoencoders were used for feature extraction, and SVM, singular value decomposition (SVD), K-means clustering, and motion analysis were used for summarization.

A consensus needs to be reached by gastroenterologists regarding the summarization depth required in clinical practice. Some experts may require the removal of duplicate and redundant frames, while others may require further reduction in reading time by removing confident normal mucosal findings. To date, as an open-choice problem, researchers have provided summarization solutions construing the required depth in their own way. Therefore, the evaluation metrics seem to be disparate rendering results that are incomparable. Many studies have used highly selective pathological data; even short clips are trimmed, rendering the quoted results less robust. Manual human-set thresholding is used as a parameter for summarization depth in most methods. However, the use of a limited dataset and the trade-off between misdetection and summarization both suggest an optimal point to be reached based on algorithmic or data-driven optimization, rather than a human-selected threshold that is vulnerable to be less optimal for a diverse nature of lesions and anomalies.

Clinicians have expressed concern about the delineation of validation methods and the study population of data used for data-driven diagnosis methods [15]. However, most datasets used appear to provide little or no such information. Therefore, these parameters are also summarized while comparing the existing methods in Table 2. These shortcomings in existing summarization solutions need to be critically addressed to make an impact on the reading time of VCE videos, indeed a dire need of hours in VCE clinical practice. Furthermore, it has been observed that existing software enhancements (based on conventional ML) have been largely unsuccessful in reducing reading time without compromising the misdetection rate [70].

4.2. Recognition and identification

Recognition is a generic term in computer vision that involves pattern recognition tasks such as classification, detection, localization, and segmentation. Any analytical approach that delineates an image visually or semantically manifests itself within the scope of image recognition. However, there is much confusion between these terminologies in the literature on visual analysis of VCE. In particular, segmentation, detection, and localization have been confounded with each other in the literature [71–78]. Classification is the task of assigning a unique label to the image. Segmentation is pixel-wise or block-based classification into various labels under study. Detection or localization is both classifying and locating the position of an object/class on the image in the form of a bounding box. CADx and CADe are terms widely used in computational medical imaging. In VCE video analysis, CADx refers to the pathological characterization of a frame while CADe refers to the localization of a pathology in the frame.

Several studies have been contributed in the CADx and CADe domains of the VCE, especially in the last decade. Table 3 presents a brief comparative list of notable contributions. Apparently, single anomaly classification tends be the favorite area explored by researchers. Active bleeding or blood content-based anomaly recognition seems to be the most explored single anomaly, primarily due to more no. of clinical cases for small bowel, and bleeding confers to be the most common indicator for SB-VCE [79–86]. Interestingly, the recognition of blood content can be a more unequivocal and visually comprehensive task from both expert and computer vision perspectives. Higher intensity levels of the red channel are easily detectable using color space histograms and other color-based features [79,81,83]; therefore, excellent performance metrics are obtained by both handcrafted features and deep learning-based feature extraction methods. Polyps and tumors, although rarely present in the small bowel as compared to other GI organs, are also widely studied for single anomaly recognition. Protruding lesions and polyps tend to be discriminated from other anomalies in terms of geometric features. However, color and textural information also appear to be discriminative among different types of protruding lesions. Indeed, the diverse appearances of polyps and various tumors among different patients is a challenging task for both experts and machine vision [87,88]. Researchers have also used frequency domain features such as wavelet transform and log Gabor filter along with other spatial features such as LBP and SUSAN edge detectors to segment the polyps as ROI [78,89,90]. The triplet loss function introduced in Laiz et al. [87] imposes an additional constraint on the learning process by forcing frames in a similar category, even if they appear visually diverse, to be represented by embedding vectors closely spaced in the feature subspace. Therefore, such a constraint would inform the learning put less focus on discriminative features within the same class. Researchers [88] attempted to achieve the same goal by using the concept of the nearest neighbor graph to model diverse image manifolds within the same class. Hence, forcing feature vectors evolve in a sparse autoencoder without much disturbance in the structure of the nearest neighbor graph. The performance metrics for both approaches conformed to the rationality introduced in their concepts. Although the use of CADe for polyps and tumors in colorectal colonoscopy has captured much attention, few methods exist for the detection of polyps in the small bowel. Nadimi

Table 2
A summary of existing methods for video summarization in VCE.

Year	Method	Dataset & Study population	Results	Validation method	Authors' discipline
2006, 2009	Given Imaging Inc. Quick view method [56,57]	100 proprietary VCE videos (single center); Given study population	75% reduction in time, 8% increase in miss-rate	Prospective	Multi-disciplinary
2010	Non-negative matrix factorization of greyscale images for shot segmentation and orthogonality for key frames selection [59]	8 non-public VCE videos for small bowel (single center); No study population	85% reduction in reading time with zero miss-rate	Retrospective	Multi-disciplinary
2010	A generative learning model based on Expectation Maximization doing epitomized summarization with semantic organization in the generated epitomes [66]	Short clips of proprietary VCE videos from a hospital (single center); No study population	90% reduction in frames, quantitative measure of miss-detection rate not performed	Retrospective	Multi-disciplinary
2011, 2013	Near-duplicate frames reduction using normalized cross-correlation, Ego-motion estimation based overlapped frames reduction due to peristalsis [60,61]	3 proprietary VCE videos (single center); No study population	52% reduction in frames, miss-detection for anomaly not performed	Retrospective	Computational
2013	Identification of scene changes based on VCE motion estimation at course and fine levels [58]	Selected short clips from a proprietary video (single center); No study population	52%–90% reduction in frames depending upon short clip region, miss-detection analysis not done	Retrospective	Multi-disciplinary
2014	Key frames are selected based on significant change in salient values computed by fusion of moments, multi-scale contrast and curvature [63]	Few VCE videos for single anomaly (Phlebectasia)- (single center); No study population	Results measured in terms of Precision, Recall and F-score. F-score = 85%	Retrospective	Computational
2014	Mobile cloud based summarization using JD divergence and Boolean series correlation for key frames extraction [64]	Few VCE videos for single anomaly (Phlebectasia)- (single center); No study population	Results measured in terms of Precision, Recall and F= score. F-score = 82%	Retrospective	Computational
2016	Siamese Neural Networks (SNN) for learning similarity based feature vectors, SVM for shot segmentation, adaptive K-means clustering for key frames selection [67]	50 proprietary VCE videos labeled in terms of similar and dissimilar frames by experts (single center); No study population	Results measured in terms of Compression Ratio (CR) and F-score. CR = 85 (avg.), F-score = 84 (avg.)	Retrospective-3 fold cross validation	Computational
2021	Deep learning based hybrid method comprising variational autoencoder based LSTM architecture with pointer network and de-redundancy mechanism providing a summary [68]	32 VCE proprietary videos (single center); No study population	Results measured in terms of Precision, Recall and F= score. F-score = 44%	Retrospective	Computational
2021	Autoencoders for feature extraction of consecutive frames, Euclidean distance based shot segmentation, key frame selection using motion analysis [69]	3 VCE videos from KID dataset, 20 proprietary VCE videos (multi-center); No study population	Results measured in terms of Precision, Recall and F= score. F-score = 92%, Compression Ratio = 84	Retrospective	Computational

et al. [91] used a faster region based convolutional neural network (RCNN) to localize colorectal and small bowel polyps using datasets from both organs, since small bowel polyps rarely occur. Therefore, data sparsity might be a reason for less work. Recently, real-time polyp segmentation for VCE and colonoscopy was demonstrated with almost the same performance as before, but with extremely low number of parameters and an enhanced frame rate for conducting real-time performance [92]. Ulcerative lesions are also diverse in each aspect of their visual appearance, including color, shape, and texture. Superlative ulcers tend to be more discriminative than deep ulcers considering they have more visual complications that may sometimes overlap with other lesion discriminative sets. Patch-based super-pixel saliency maps and second-glance refinement were employed for classification and detection, respectively, in Yuan et al. [93] and Wang et al. [94]. Similarly, single anomaly recognition for angioectasia, hookworm, and celiac have been performed in He et al. [73],Leenhardt et al. [75],Tsuboi et al. [95], and Zhou et al. [96], respectively.

Learning to discriminate between a single type of anomaly and normal mucosa can be a relatively easier task (binary classification problem). However, such a confined approach is rendered unpragmatic in clinical settings. Multi-lesion recognition in VCE is a complex task, considering low resolution and improper illumination conditions coupled with untamed motion of the camera might present one type of lesion or anomaly in diverse visual representations, thereby reducing powerful discriminative features. Some researchers have contributed to multi-anomaly recognition tasks. The number of anomalies recognized by these CAD systems

appears to be a matter of choice among researchers, and usually varies from 2 to 7. Researchers in Sharif et al. [74] and Khan et al. [97] classified two types of lesions (ulcer and bleeding) using a transfer learning approach with VGG-16 and VGG-19 as the backbone. Transfer learning has been widely applied to VCE recognition tasks over the last five years. Some researchers, mostly from the medical domain, have used transfer learning with fine-tuning according to the problem at hand [72,76,84,85,95,98–102]. Those from a computational background have fused transfer learning with more customized methodologies to achieve better results [74,77,87,91,94,97,103–106]. Interestingly, a multi-task learning (MTL) approach was adopted in Vats et al. [107] to primarily address two challenges in a VCE multi-anomaly recognition task: possible similarities in visual characteristics across different conditions and their different levels of severity; simultaneous existence of non-pathological, visually prevalent similarities among various frames. In the MTL, additional self-chosen tasks are defined to supplement domain-specific learning by sharing information across several related tasks. However, the cardinal learning objective is still weighed more in the overall objective function. Comparison of results between MTL and single-task learning (STL) for classification of vascular and inflammatory lesions and normal tissue suggests that this is an efficient approach. However, the results must be verified in a more robust manner. In a multiple lesion classification attempt [105], few-shot learning leveraging on deep metric learning was presented to address the data sparsity problem in a multiple-lesion detection system for VCE. Lumping different anomalies into one category named as significant and the remaining normal and variants into another category named as non-significant for a binary classification using Inception-ResNet-v2 seems efficacious in highlighting multiple anomalies without classifying the anomaly type [99]. Ding et al. in a large multicenter dataset study so far on SB-VCE to detect multiple anomalies, presented ResNet-152 by training for binary classification into abnormal (containing 10 types of abnormalities) and normal [108]. Both studies have been validated in a more pragmatic way by creating two types of reading groups: the conventional reading group and AI-assisted reading group. The AI-assisted reading group showed much better performance in terms of lesion detection rate and reading time. However, classification into a specific lesion or abnormality name at the inference stage is omitted in these works.

Existing research methodologies for the task of VCE image/frame recognition do significant work; however, some technicalities are identified that need to be addressed before transforming into a reliable clinical solution. In the realm of machine learning, both training and validation require representative datasets from independent sources. Highly selective data for training render the model a less generalizable solution [73,77,83,88]. In particular, methods where handcrafted features are generated must be extensively checked for all types of diversity in lesion shape, color, and texture [81,89,90]. Dataset availability for the performance benchmarking of different techniques must be addressed. Regulatory bodies essentially require the study population in a medical setting to analyze possible biases and imbalances in data representation. Most studies have not provided the study population for their datasets, except for a few [75,96,100,101,109,110]. All existing studies have been validated via retrospective validation; however, prospective validation entails more confidence in the applicability of the proposed solution, and also offers the real scenario to perform in Soffer et al. [15]. Even in retrospective validation, the random split between the training and validation sets is based on the frame level in most studies except [59,73,76,90,94,96,108]; however, it should be on a patient- or video-level basis to ensure full independence between the training and validation frames. The lack of standard metrics for the evaluation of proposed methods by researchers increases the challenges in comparing the results. In summary, recognition tasks are mostly applied for single anomaly detection, and a potential solution may be achieved in the future for large datacenter-based multiple anomaly recognition covering all possible anomalous conditions under a single task by addressing critical observations.

4.3. Localization and active locomotion

It is crucial to know the location of the capsule inside the body to locate anomalies and lesions. The accurate location of an anomaly could be helpful in further targeted investigational procedures or for therapeutic purposes. Merging localization with 3D mapping of the capsule's followed path, also called simultaneous localization and mapping (SLAM), is a significant task in machine vision for robotics. Similarly, active locomotion could potentially turn a passive and untamed video capsule into a controlled robot. This motion control may potentially harness the full spectrum of benefits of the CE. Currently, the capsule moves untamed under gravity and peristaltic movements, which are highly dependent on the patients' gastric activity and retention levels. Therefore, the motion profile is non-uniform, and sometimes longer gastric retention may result in incomplete capture owing to limited battery life. In this sub-section, we present a brief overview of several computational approaches for localization and active locomotion. Several localization and active locomotion approaches have been presented by researchers using modalities such as ultrasonic imaging, magnetic resonance imaging (MRI), positron emission tomography (PET), fluoroscopy, radio frequency (RF), and magnetic-field-based techniques [114–118]. However, here we present only the machine learning-based computational approaches for these two tasks. In the context of computer vision, visual odometry is a well-defined task for motion and pose estimation of the capsule. Classical visual odometry steps include feature extraction, feature tracking, rigid body motion estimation, and joint adjustment. The inputs to such a system are images and the output is a 6-DoF pose. Turan et al. [119] proposed a novel deep-learning-based visual odometry method for capsule endoscopy. They employed a recurrent CNN (Recurrent CNN) architecture to model subtle and complicated motion dynamics across endoscopic frames. One significant challenge in the localization of the capsule is the validation procedure. Knowing the ground truth for a pose during a real-time CE procedure is complex. Investigative clinical procedures, such as planar X-ray imaging and ultrasound, cannot be extensively used to verify capsule estimated pose owing to their associated costs and health risks. The researchers in this work used a real pig stomach and synthetic human simulator dataset with a 6-DoF pose as the ground truth. A normalized depth image was created from RGB images, and depth images for consecutive frames stacked together as tensors were fed into the inception CNN architecture to form the feature vector. The LSTM-based RNN uses this sequence of feature vectors to estimate 6-DoF pose of the capsule. A performance comparison with state-of-the-art SLAM methods, such as large-scale direct monocular (LSD) SLAM [120] and oriented fast and rotated brief (ORB) SLAM [121], shows better translation and rotation error profiles. A hybrid approach was adopted in Bao et al. [122] by merging visual odometry with RF-based localization. For pose estimation using visual images, they employed VO steps such as feature point detection using the ASIFT descriptor, image unrolling to detect motion, and estimation of speed and direction of motion. Visual motion tracking is prone to cumulative estimation errors that drift the estimated pose over time. RF localization is absolute in its estimation without

Table 3
A summary of existing recognition methods for Video Capsule Endoscopy.

Year	Task	Method	Dataset & Study population	Results	Validation method	Authors' discipline
2011	Classification, Segmentation of bleeding [79]	Segmentation of bleeding using color texture features in RGB and HIS, Probabilistic Neural Network (PNN) for pixel-level classification	Selective images from 150 proprietary VCE videos (single center); No study population	Pixel level: Sensitivity-87, Specificity-85; Image level: Sensitivity-93 Specificity-86	Retrospective with frame level cross validation	Computational
2011	Classification, Segmentation into polyps or ulcer [89]	Log Gabor filter, SUSAN edge detectors, color, and texture features to segment ROIs, SVM for classification into polyp or ulcer	Highly selective data, short clips of 50–60 frames for ulcer and polyps (single center); No study population	Sensitivity: 100, Specificity: 75	Retrospective with frame level cross validation	Computational
2012	Classification into tumors and normal [90]	Uniform LBP and wavelet transform based features, SVM-SFFS and SVM-RFE for feature selection and classification	Selective frames of tumors from 10 proprietary videos (single center); No study population	Sensitivity: 87, Specificity: 92	Retrospective, Patient level 10-fold cross validation	Computational
2015	Classification into various GI organs [111]	Deep CNN based architecture with 3 convolutional layers	Selective frames of 30 prop. videos (single center); No study population	Accuracy: 95.5	Retrospective with frame level cross validation	Computational
2015	Classification of ulcers vs. normal [93]	Fusion of color and textural features from multi-level super-pixel groups into saliency maps, Saliency based locality constrained linear coding for classification	Proprietary dataset containing 130 ulcer and 130 normal frames (single center); No study population	Sensitivity: 94, Specificity: 91	Retrospective with frame level 5-fold cross validation	Computational
2016	Classification into motility conditions [112]	Deep CNN with 5 layers and additional channels of Laplacian and Hessian merged in two different ways	Selective frames from 50 proprietary videos (single center); No study population	Accuracy (best): 96	Retrospective with frame level cross validation	Multi-disciplinary
2016	Classification into bleeding and normal [80]	Deep CNN with 8 layers for features extraction, SVM as a classifier	10,000 selective frames from proprietary videos (single center)	F1 score: 99.5, Precision: 99.9, Recall: 99.2	Retrospective with frame level cross validation	Computational
2016	Segmentation of bleeding ROIs, Classification into bleeding and normal [81]	Color channel mixing and visual contrast-based saliency maps for segmentation, ROI color channel features for mapping frames into words of color histogram, SVM and KNN for classification	2400 selective frames of bleeding and normal from 10 proprietary videos (single center); No study population	Sensitivity: 92, Specificity: 97	Retrospective with frame level cross validation	Computational
2017	Classification of celiac and normal, disease level prediction in terms of probability [96]	GoogLeNet is used to train initially on most representative frames for achieving optimal gradient profile, later rest of frames are used	Selective frames from 21 proprietary videos (single center); Given study population	Disease level sensitivity and specificity: 100	Retrospective with patient level 7-fold cross validation	Multi-disciplinary
2017	Classification into polyps vs. other motility conditions [88]	Stacked sparse autoencoder with image manifold constraint in the cost function to cater for visual diversity in VCE frames	Selective 4000 frames from 35 proprietary videos (single center); No study population	Accuracy: 98	Retrospective with frame level cross validation	Computational
2017	Classification into hemorrhage and normal [82]	4-way data augmentation: rotation, illumination, blurriness, poison noise to cope with data sparsity and diverse visual nature, Transfer learning models for classification	12000 selective frames from proprietary videos (single center); No study population	F-score: above 95 for all four models (Le-Net, Alex-Net, GoogLeNet, VGG); data augmentation improved F-score	Retrospective with frame level cross validation	Computational
2017	Classification, segmentation of active blood normal [83]	Color histogram features based classification into positive or negative, Deep CNNs for segmentation of active blood.	Highly selective 300 frames form proprietary videos of 12 patients (single center); No study population	Mean IU: 77.5 Mean accuracy: 87	Retrospective with frame level cross validation	Computational
2018	Classification of ulcer and erosion vs. normal in 2 independent models [72]	Alex-Net based transfer learning approach for binary classification into ulcer or erosion vs. normal	Selective images from 144 proprietary videos (single center); No study population	Average results Accuracy: 95 Sensitivity: 95.2, Specificity: 95.7	Retrospective with frame level cross validation	Multi-disciplinary

(*continued on next page*)

Table 3 (*continued*).

Year	Task	Method	Dataset & Study population	Results	Validation method	Authors' discipline
2018	Classification into anomalous and normal, anomaly region detection [71]	Custom Deep CNN architecture for classification, Deep features based salient point detection for localization of anomaly	KID dataset 2, Gastroscopy challenge dataset (non-VCE); No study population	Accuracy: 90 Sensitivity: 92, Specificity: 87	Retrospective with patient level cross validation	Computational
2019	Classification + detection of ulcers [94]	Primary detection using Retina-Net, second glance patch and image level refinement built on ResNet-18 and ResNet-34 backbone.	Selective frames from proprietary dataset of 1504 patients (multi-center); No study population	Accuracy: 90 Sensitivity: 89.7 Specificity: 90.5	Retrospective with patient level cross validation and test	Computational
2019	Classification + detection of angio ectasia [95]	Deeps CNN based transfer learning using Single Shot Multi-Box Detector	Selective images 189 proprietary videos (single center); No study population	Sensitivity: 98.8 Specificity: 98.4	Retrospective with frame level validation	Medicine
2019	Classification into bleeding, normal and ulcers [74]	Fusion of VGG-16, VGG-19 features and geometric features extracted via handcrafted segmentation technique, KNN for classification	Selective frames from a proprietary collection of 10 videos (single center); No study population	Accuracy: 99 Sensitivity and Specificity: 100	Retrospective with frame level 10-fold cross validation	Computational
2019	Classification, segmentation of angioectasia [75]	Deep CNN based customized architecture for pixel level classification	Selective 6360 images from CAD-CAP dataset (multi-center); Given study population	Sensitivity:100 Specificity: 96	Retrospective with frame level split test set	Multi-disciplinary
2019	Detection, classification of erosions and ulcers [76]	Transfer learning approach using Single Shot Multi-Box Detector	15,800 frames from 180 proprietary videos (single center); No study population	Accuracy: 90.8 Sensitivity: 88.2 Specificity: 90.9	Retrospective with patient level split test set	Multi-disciplinary
2019	Classification into normal and blood content [84]	Transfer learning approach using ResNet-50	Selective frames from 66 proprietary videos (single center); Partially provided	Accuracy: 99 Sensitivity: 96.6 Specificity: 99.9	Retrospective with patient level split test set	Multi-disciplinary
2019	Classification into ulcer and normal [77]	Transfer learning approach using GoogLeNet, AlexNet	1875 images from proprietary dataset (single center); No study population	Accuracy, Specificity, Sensitivity: 100	Retrospective with frame level split test set	Computational
2019	Multi-anomaly classification into abnormal and normal [113]	ResNet based transfer learning (only multi-center and multiple anomaly detection study in VCE domain)	Large multi-center proprietary dataset 6970 patients; No study population	Sensitivity: 99 Specificity: 100	Retrospective with patient level separate validation set of 5000 cases	Multi-disciplinary
2020	Classification into 7 lesion types and normal [98]	2 ResNet-34 and faster RCNN based framework	Selective frames from proprietary 797 videos (multi-center); No study population	AUC (for all anomalies) = 84	Retrospective frame level cross validation	Medicine
2020	Detection of small intestine lesions [103]	YOLO-v3 based lesion localization and classification	3120 lesion and normal images from proprietary videos (single center); No study population	Mean Average Precision (mAP): 93 fps: 21	Retrospective frame level cross validation and test set	Computational
2020	Classification into significant and non-significant frames [99]	Inception-ResNet-v2 transfer learning approach with fine tuning (various lesions and anomalies are lumped under one class 'significant')	Selective images from proprietary 139 videos (single center); No study population	Accuracy: 98.3 Sensitivity: 96 Specificity: 99.5 AUC: 99.8	Retrospective fame level cross validation, validated also on a patient level different set in AI vs. expert mode.	Multi-disciplinary
2020	Classification into normal mucosa and mucosal ulcers [100]	Xception CNN based transfer learning approach	Selective frames from proprietary 49 videos (single center); Given study population	Accuracy: 95.7 Sensitivity: 94.5 Specificity: 97	Retrospective frame level split 5-fold cross validation, one to many patient level validation	Medicine

(*continued on next page*)

depending on previous estimations; therefore, the accumulative error is absent. A Kalman filter was used to estimate the pose of the capsule using results from visual odometry, and the feedback loop involved RF measurements to correct the pose estimations. When compared with existing RF-based localization systems, the average localization error reportedly reduced from 6.8 cm to less than 2.3 cm. Active locomotion can be implemented in two ways: internal locomotion using propellers or paddles attached to the capsule's shell, as in a robot, or external locomotion using externally applied stimuli such as a magnetic force. Internal locomotion is problematic owing to the power constraints and unstable movements. Therefore, most magnetically controlled external locomotion methods are considered feasible and have been extensively studied. However, such active locomotive systems are

Table 3 (*continued*).

Year	Task	Method	Dataset & Study population	Results	Validation method	Authors' discipline
2020	Classification into hemorrhagic, ulcerative and normal [104]	Modified VGG-Net for classification, Grad-CAM used to visualize class activation maps	Selective frames from proprietary 526 videos (multi-center); No study population	Accuracy: 96.8 Sensitivity: 97.4 Specificity: 98	Retrospective validation on a different set of 162 videos from another center	Multi-disciplinary
2020	Classification of ulcer, bleeding and normal [97]	Fusion of VGG16 and GLDM features, PSD grand mean-based feature selection, cubic kernel SVM for classification	6000 selective frames from proprietary small dataset (single center); No study population	Accuracy: 98.3 F1-score: 98.4	Retrospective frame level 10-fold cross validation	Computational
2021	Classification into different lesions with associated hemorrhagic potential [102]	Xception based transfer learning approach	Selective frames from proprietary 5793 videos (multi-center); No study population	Accuracy: 99 Sensitivity: 88.4 Specificity: 99.2	Retrospective frame level cross validation	Medicine
2021	Polyp segmentation (real-time) for VCE and colonoscopy [92]	Novel Nano-Net architecture based on encoder–decoder framework, MobileNetv2 used as encoder, modified Residual block used as decoder.	Fusion of Kvasir-capsule (polyps only), Kvasir-seg. and other publicly available colonoscopy polyps datasets; No study population	Matched performance with SOTA methods, at increased fps and less no. of parameters	Retrospective frame level cross validation	Multi-disciplinary
2021	Classification of blood vs. no blood [85]	Xception model-based transfer learning with fine tuning	22095 selective frames from proprietary dataset (single center); No study population	Accuracy: 98.5 Sensitivity: 98.6 Specificity: 98.9	Retrospective frame level cross validation	Medicine
2021	Classification of frames into 5 multiple lesions and normal [109]	Framework comprising 3 SSD network and one ResNet-50 for classification of various lesions	Selective images from proprietary multicenter videos; No study population	Average detection rate for all lesions: 98 (better than quick view)	Retrospective patient level cross validation	Multi-disciplinary
2021	Classification of frames into bleeding vs. normal [86]	A cascaded Mobile-net and custom deep CNN based model	Selective frames from 33 proprietary videos (single center); No study population	Accuracy: 99.3 F1-score: 99.7	Retrospective frame level split validation	Computational
2021	Binary classification between significant and non-significant frame [110]	Inception-ResNet-v2 based transfer learning approach	400k frames selected from 84 proprietary videos (multi-center); Given study population	Cross-validation accuracy: 98 External validation accuracy: 85.7 AUC: 92.2	Retrospective frame level cross validation, validation also done on data from other center	Multi-disciplinary
2021	Classification into four types of lesion [105]	Deep metric based few shots learning incorporated into base models of AlexNet, VGG, ResNet	5360 frames selected from 52 proprietary videos (single center); No study population	Best accuracy: 90.8 F-score: 91 (Alex-Net)	Retrospective frame level cross validation	Computational
2022	Attention augmented classification into ulcer, blood, polyp and normal [106]	ResNet-50 as backbone, lesion self-attention (local and global) maps fused with original frame to elevate lesions for classification network	Selective frames from Kvasir capsule merged with bleeding dataset. (multi-center); No study population	Average accuracy: 95.1 (Kvasir), 94.7 (bleeding detection data)	Retrospective frame level 4-fold cross validation and test set	Computational

still in the validation phase, and no such commercial systems are currently available.

4.4. State-of-the-art in clinical practice

In this sub-section, we provide an overview of the state-of-the-art clinical systems and software features of various leading manufacturers of capsule endoscopy. Medtronic's latest system with the name of "PillCam ™ SB 3 Capsule" [123] along with paraphernalia that comprises a sensor belt, sensor array, and a recorder (ver. 3) is more efficient and effective than older versions. A major improvement in hardware is the introduction of adaptive frame rate technology. The frame rate automatically adjusts itself from 2 fps to 6 fps depending on the motility conditions, thereby facilitating efficient capturing while ameliorating redundancy and mucosal coverage. In addition, the wide-angle view, image quality, and battery charge capacity of the capsule have been improved from previous versions. The reading software has been improved with the name of "PillCam™ Software V9 Update". The reading time is claimed to be 10% faster than the previous version. A new "Top 100" feature has been introduced listing the top 100 most clinically relevant frames to assist in identification of pathologies in a lesser time. A quick-view mode enables a rapid study preview to cover videos in a few minutes. A 2-D simplistic GI map was introduced to view the progress of the capsule in real time. Accordingly, the necessary audio and visual instructions are passed to the patient to help progress the capsule into the forward stages. Similarly, Olympus introduced its latest VCE system with the name of "Endocapsule 10" [124]. The omni-mode has been claimed to reduce the reading time by up to 64% without compromising the diagnostic outcome. The omni mode detects the similarity in frames even if the angle of capture is changed across two frames belonging to the same scene. Improvements in image quality, halation, noise reduction, and adjustment of brightness levels to balance across diverse brightness conditions have been reported. The angle of view increased from

145° to 160°; likewise, the battery lifetime increased from 8 to 12 h, which increased the observation time by 50%. A real-time 3D tracking function on the screen allows the patient to view the progress of the capsule movement along the GI tract. The same visualization is available on a per-frame basis to localize the lesion in the small intestine. In nutshell, reading time has been significantly reduced by AI (deep learning) based omni mode and quick-view mode in Olympus and Medtronic systems respectively [70,125]. AI-based efficient detection of multiple anomalies has yet to be introduced in these advanced endo-capsule systems.

5. Challenges and opportunities

5.1. Intricate nature of VCE images

Images acquired by a capsule – by virtue of nature of the process by which they are captured – are indeed intricate over manifold aspects. Untamed motion of capsule under peristalsis and gravity certainly violates the benchmark rules for capturing effective and informative shots. Appropriate distance from target scene, stillness of camera for focusing on a specific point, angle of capture to avoid possible light reflections, zoom control for a desired scene are some of the key standards being complied in any visual scene capturing for analysis, especially conventional endoscopic procedure. An expert has full control over maneuvering camera inside the lumen to delineate the suspected areas in conventional endoscopy. Conversely, nearly wild motion profile of capturing device, i.e., capsule, posits certain challenges pertinent to image analysis and interpretation in the case of VCE. Furthermore, air is also not inflated in case of VCE conducing to poor luminal volume for capturing effective shots particularly for small bowel—already a narrowly convoluted structure. Miniaturization of optics and electronics into a pill sized capsule somehow affects the overall image quality in terms of resolution and noise, which is unlikely in conventional endoscopic probe. These procedural differences in VCE create subtleties and intricacies among images captured. Visual similarity among frames from different categories is the most pronounced subtlety caused. Diverse capturing scenarios for various luminal structures whether normal, normal variant, mildly anomalous or severely anomalous tend to lessen the visual discriminative power among various classes. Normal frames belonging to anatomical landmarks may visually overlap with abnormal classes. For example, Pylorus may appear like tumor or mass captured from the duodenal side. Similarly, ampulla and ileocecal valve could be mistaken for a polyp owing to visual similarities in some frames. Lymphatic structures such as lymphectasia and lymphatics cysts are vulnerable to be mistakenly construed as more pathological appearances such as lymphangioectasias. Tumors and polyps may be captured in a confined side view or closed view creating confusion with regular mucosal bulges and rounded folds. Red spot and angioectasia are similar and difficult to differentiate. Similarly, erosion and ulceration are visually similar in respective anomalous regions of some frames. Fig. 7 shows some of such similar pairs from different categories. An anomalous frame comprises active anomaly region and background region, which itself could be very diverse in visual appearance. Frames where anomaly is visually well pronounced relative to the background scene are somehow obliging to be detected. However, problems arise in such frames when the anomalous region tends to match or overlap the anomaly signature for some other class. However, there exist other types of frames where anomaly may not be much pronounced visually, and hence, background scene becomes a dominant player in learning class activations. In such scenarios, multi-category classification becomes vulnerable to in-accurate results as background scene is quite likely to be nearer to other landmarks classes or normal/variant mucosal categories such as in Fig. 7 (case g).

Visual diversity in the background of different anomalous frames appears to be a challenge that must be addressed, particularly for frames with diffused or mild manifestations of anomalies. One such case of erosion is shown in Fig. 8. The images were adapted from an open access dataset [49].

The subtle manifestation of lesions against highly pronounced diverse backgrounds in frames may possibly confound the neurons in learning the discriminative features for each category in a multiclass classification task. For example, image on top left corner would possibly be wrongly classified in the ‘bubbles’ or ‘obscure’ categories if such classes exist in labels. Else, the dominating bubbles or other backgrounds that manifest themselves in other classes would interfere with the process of learning class activation patterns or could slow the learning process owing to the effort being wasted on learning the connection of the image-dominant background with a peculiar type of anomaly.

Supervised learning with anomaly annotation as a bounding box or pixel-wise segmentation can help cope with learning discriminative features of mildly manifested lesions as anomaly location could be leveraged by ignoring the visually dominant diverse backgrounds, and learning could focus only on lesion-specific areas to avoid misleading backgrounds. However, inter-class visual similarity is a problem that must be addressed meticulously while designing a learning system. Based on a comprehensive literature review and comparative analysis of state-of-the-art systems, we suggest constructing a class labeling hierarchy to cater for visual similarities among various lesions without significantly perturbing the pathological essence of frame-level labels, as shown in Fig. 3. This is because the hierarchy of their clinical relevance may arguably help in creating better margins to discriminate among different classes. Designing a more robust system capable of forcibly learning only discriminative features in such a confounding multiclass dataset is another future direction and challenge. As discussed earlier, some researchers attempted to address the challenges of visual similarities and diversities in Vieira et al. [78],Li et al. [82] and Yuan and Meng [88] by proposing an additional image-manifold constraint in the cost function, extended data augmentation, and ensemble learning. However, in a multiclass prospective setting, the results must be validated to ensure the rationality of the proposed methods.

5.2. Datasets related problems

The availability of VCE datasets as open-access academics is a significant problem. Unfortunately, many publicly available datasets have now been termed proprietary datasets and are not available to researchers working in this domain [52–54]. Most studies have utilized proprietary datasets collected and compiled by the authors themselves and have never been made accessible to the research community. This may mark a big question to the rationality of the comparisons in terms of performance metrics made in many articles when the foundations, that is, datasets, have been different across these experiments. Data sparsity is another significant challenge for the VCE datasets. Not every sample (patient) contributed equally to the different frame-level categories, specifically for anomalies. Some anomalies are less likely to occur or manifest in small-bowel VCE, such as polyps, tumors, hookworms, and sparsity in the number of available frames for such anomalies engenders a typical problem in machine learning known as class imbalance. Class imbalance precludes achieving better results on unseen data because of the inherent bias introduced. In multiple-lesion detection, class imbalance must be addressed efficiently. Extended data augmentation and sampling techniques have been frequently employed to cater to imbalances in classes [16,82,86].

Fig. 7. Some examples of mimicker images from different classes having visual similarities: (a) Polyp, Pylorus; (b) Lymphangiectasia, Lymphectasia; (c) Red spot/dot, Angiectasia; (d) Intestinal lymphectasias, Ulceration; (e) Normal mucosal bulge, Sub-mucosal carcinoid; (f) Angioectasia, Erythema; (g) Ulcer, Ileocecal valve; (h) Erosion, Ulcer.
Source: Images taken from [38,49].

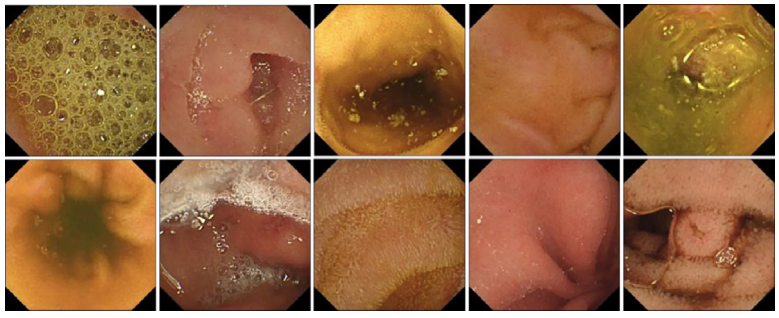

Fig. 8. Example of anomalous frames with mild manifestation of anomaly, yet diverse background scenes.
Source: Erosion frames adapted from [49].

Many dissensions exist in frame-level label terminology across various datasets and research studies, as discussed earlier. Therefore, an attempt to merge all the data from various centers requires extra effort to either label them again according to a common standard labeling terminology for small bowel VCE or devise a label merging scheme according to pathological closeness. Therefore, it is necessary to follow an international consensus developed on frame-level labeling terminology introduced in Leenhardt et al. [39,40] and Leenhardt et al. [41] in the future. Bounding boxes around anomaly or pixel-level lesion highlighting can potentially identify vague anomalies and lesions accurately; however, generating consensual ground truth for them is a very tedious and time-consuming process, particularly for VCE, where either the image quality is relatively compromised or the view is not conducive. Interactive and user-friendly frameworks should be designed to make this tedious and boring task interesting and less time-consuming.

While labeling VCE frames, especially for anatomical landmarks, extra care is required when labeling only representative frames where specific landmark organs are vividly present. In a video, many consecutive frames in the form of a shot could be related to that anatomical landmark; however, it is quite likely that only a few frames in that shot may contain actual representative images of the landmark. This practice can potentially mitigate the visual similarity challenge among various categories, particularly for anatomical landmarks. Some proprietary datasets used in research studies lack study population information and inclusion and exclusion criteria, which are considered significantly important for understanding the limitations and biases of studies based on such datasets. In summary, large multicenter datasets labeled and annotated according to the consensual frame-level findings standard must be created, keeping in mind the possible visual vantage point for inclusion or exclusion of a frame in a particular category of the training set.

The scarcity of VCE datasets for research purposes has attracted significant attention for high-fidelity synthetic data generation. Fabricating synthetic data by allowing a model to learn all the underlying intricacies (probabilistic distributions, nonlinearities, and noise effects) from real data, without imitating patient privacy aspects from real data, confers a handsome solution to cope with data paucity as well as ethical and legal bindings in healthcare. Circumventing privacy concerns while simultaneously maintaining high fidelity is a challenge. GANs are extensively employed in other computer vision domains, where less imaging data is available and the results are quite propitious [126,127]. In medical imaging, there have been a few recent attempts to apply GAN to overcome data scarcity and class imbalance [128–131]. The results are arguable, and GAN needs to undergo further evaluation for its usefulness in medical imaging, particularly capsule endoscopy [128,132]. Additionally, synthetic data generation raises its own privacy and security vulnerabilities, especially in medical imaging, and poses significant challenges to healthcare administrative and legal policies [133].

5.3. Reliability and explainability of machine learning based diagnostics systems

The term ‘reliability’ attains phenomenal significance for machine-learning-based diagnostics systems in healthcare. Relying on machines or the mere assistance for critical decisions pertinent to precious lives implies high standards of performance and quality. Although the standards of quality and performance have a much wider spectrum, only machine-learning-related quality and performance are relevant here. Machine-learning-based diagnostics or decision support systems are mostly data-driven; therefore, the quality of data on which a model is trained should be very high in terms of representation power, bias, population diversity, and consensually generated ground truths. Failure to comply with data quality standards may provide acceptable performance on retrospectively collected validation splits; however, such a system might not perform well in a real-world heterogeneous prospective setting. Poor quality data imparts multi-faceted bias, leading to less generalizable and less robust, and in turn, an unreliable system. Many of the VCE CAD systems in research studies lack quality-related data information. Apart from the data, the explainability of the model itself plays a vital role in the overall system reliability. In particular, deep learning models are being prolifically employed as black boxes without tapping into the explainability factor. Explainable AI (XAI) is a popular research area owing to its significance in creating reliable systems for critical applications such as healthcare and self-driving cars [134]. Lack of explainability may further bolster ethical and legal issues that impede the clinical use of such innovative computer-aided detection systems. Tapping into several deep learning models applied in both clinical and non-clinical settings has recently revealed telltale observations regarding the reliability of such performances [135,136]. Clinicians, as end users, only accept and approve these “black-box” systems once they represent a high degree of explainability, interpretability, transparency, and traceability.

XAI refers to AI systems with an insightful rationale for their decision-making processes. Behavioral outputs are well-reasoned, understandable, and explorable to depict potential pros and cons; therefore, the systems are more trustworthy and transparent to their end users [137,138]. In the medical domain, only generally perceived clinically relevant features should be the foundation basis for important decision-making or analysis. Conventional machine learning-based systems are usually more explainable and interpretable compared to modern deep-learning-based systems; however, the performance of DL-based systems have proven to be better than the former ones. Deep learning-based decisions in healthcare are easily susceptible to depending on unwanted odd factors such as bias in certain demographic parameters of training data rather than clinically relevant contextual reasons. An XAI system is believed to possess intrinsic characteristics such as explainability, interpretability, transparency, justifiability, and contestability [139–141]. Several approaches have been adopted by researchers to employ XAI in healthcare, such as XAI by dimensionality reduction (using PCA, ICA, etc. to simplify features enhancing interpretability) [142–144], feature importance (finding correlations and domain relevant reasons between features and outcomes) [145–147], attention-based visualizations (attention-based saliency maps, CAM, Grad-CAM, etc.) [148–150], and surrogate representations (LIME) [151, 152]. In the VCE context, CAD-based analysis can be validated by bridging or at least explaining the gap between the outcome of explainability methods such as class activation maps, activation maximization, reverse engineering CNNs, and clinical relevance established by medical experts in several case studies [140]. Adopting a modular approach, such as the framework proposed in this work (considering frame level and disease level tasks as separate), may offer more convenience in applying XAI methods to CADx for capsule endoscopy. Additionally, incorporating the explainability factor into the objective function of a learnable model during the training process may enable the design of XAI-aware systems with enhanced interpretability and explainability for end-users [141].

Traceability and reproducibility can hamper the adoption of CAD systems in clinical practice. Tracing back the outcome through all transformations to the input data with clarity regarding each processing step along the pipeline is traceability [153]. Reproducibility is broader in spectrum-encompassing methods,

results, and inference-related reproducibility. Together, the traceability and reproducibility resolve the overall transparency of the proposed system [154,155]. The transparency of a CAD system affects the confidence level of reliance on the system in a clinical setting. Domain experts should be aware of the limitations, artifacts, working principles, strengths, and weaknesses of computer-based assistive systems so that the extent of reliance on the CAD system can be well formulated as a policy to avoid any mishaps. In summary, the reliability of an AI-based VCE analysis system is highly dependent on data quality, the inherent robustness of the model, and the extent of the system's closeness to an XAI system (explainability, interpretability, transparency, and justifiability).

5.4. Ethical and legal issues

AI-assisted capsule endoscopy faces ethical and legal challenges similar to those confronted by other AI systems in healthcare and medical imaging. In the realm of ethics in AI, four factors are deemed significant by researchers across the domain: informed consent, data privacy and security, transparency, and algorithmic fairness [156,157]. Informed consent, which is the most immediate issue engendered by AI integration in healthcare, has not received due attention. The intricacies around black-box AI cause many concerns regarding possible data or algorithmic bias risks. The extent to which a patient must be informed of such complexities is a primary ethical and legislative concern. However, from a clinician's perspective, answering these questions about risk factors is challenging. Several AI-based systems are dependent on patient data. For example, the current regulations regarding medical data require informed consent to delineate the purpose of use [158]. However, much of AI today (deep learning and unsupervised learning) reveals certain new biomarkers or has been used for new tasks not even conceived at the time of retrospectively collected data. Hence, legal bindings confine the true potential of AI to leverage. Concerns regarding data privacy, security, and consent for usage from patients need to be considered and addressed by both communities, that is, AI and legal experts [159]. The transparency of an AI system depends on its accessibility, comprehensibility, and explainability to the end user. Be it a diagnostic system or decision support system, the extent to which a patient and a physician should both know about the details or explainability of such a prospective system is yet to be formulated by regulatory bodies. Therefore, some critical questions need to be contemplated and explained purposefully, for example, whether the patient just requires knowing external agent-level information regarding AI-based systems or some details explaining the algorithm, limitations, risk, and level of transparency associated. AI here is not used as a mere endoscopic or CT scan device, which are only the imaging modalities. The reading and interpreting of results from these modalities are dependent on the intelligence and experience of a human expert, that is, a doctor, who is chosen by the patients. Hence, AI-based CAD might require a more critical evaluation of the extent of transparency and the level of physician dependency on them by legal and ethical experts. The explainability of AI-based healthcare systems is deeply rooted in connections with ethical and legal concerns. The consistently evolving and widening applications of AI in the medical domain seek germane adaptations in law and regulations to help such systems deliver their impact [160,161]. Algorithmic fairness and bias are highly related to the level of transparency and explainability. These factors can be the most dominant objections raised by the legislative bodies. Several examples exist in the literature where AI-based algorithmic biases cause injustice or inaccurate diagnosis based on ethnic origins, gender, skin color, age, or other disabilities [162–164]. Liability is another major legislative issue for AI-based systems, particularly in healthcare. Although high-performance AI systems are vulnerable to failure under certain unseen or intentionally perturbed circumstances, no one can be held accountable considering liability boundaries seem quite equivocal. However, health care governance systems attribute great significance to the principles of liability and accountability. Data protection and cybersecurity are also important concerns in addressing legal issues related to AI in healthcare. Hostile forces can manipulate data to introduce bias in AI-based decisions or to misrepresent a patient's health record for their own benefit. Much legal work has been attributed to cybersecurity and data protection in general and applies to the healthcare domain as well [165]. In their discussion paper published recently [166], the Food and Drug Administration (FDA) discussed and invited feedback from domain experts on proposed modifications in the regulatory framework for incorporating AI-based software in medical systems. While it does mention an appropriate level of transparency in the output and algorithm, the detailed levels of ethical and legal aspects and the precise extent of explainability, transparency, and liability for acceptance have not been explored. In the future, a precise and well-defined regulatory framework may be expected from legislative bodies such as the FDA and Medical Devices Regulation (MDR).

6. Conclusion

Capsule endoscopy has been proven to be the first-line gold standard for diagnosing small bowel abnormalities. However, analyzing lengthy videos for subtle anomalies among several redundant frames by a human expert can result in a high misdetection rate. AI, particularly deep CNNs, promises to solve this problem, saving both time and the misdetection rate. In this study, we developed a prospective hierarchy of tasks for analyzing VCE videos using a machine learning-based system in the same manner, where a gastroenterologist reaches the final disease-level conclusion by analyzing raw frames. We propose a taxonomy of frame-level findings to conform to both pathological and visual bases. Applying machine learning in such a taxonomical manner for classification of frames may potentially generate better classification accuracies at the lesion level for less distinct and subtle findings under the same lesion level category. To some extent, it may also ameliorate the mimicking nature of VCE frames for various lesions. Mapping frame-level findings to disease-level diagnosis along with other inputs of frame label prediction timeline, patient history, and meta-data (gender, ethnicity, and age) in the light of clinical relevance surveys could be the footprint approach for a prospective end-to-end holistic solution.

Advancements in computer vision, particularly deep CNNs, have demonstrated remarkable results for video analysis in capsule endoscopy. Despite the outstanding performance metrics cited in the research studies; the clinical implementation of these state-of-the-art methods has not yet been implemented to deliver true potential. Furthermore, by incorporating the observations made recently by clinical experts, we contemplated the advantages and disadvantages of existing machine-learning-based methods, highlighting some shortcomings or overlooked areas in this study. Several studies are prone to selective approaches with highly narrowed tasks, such as classification or detection of a single anomaly. Moreover, the validation strategies differ among the methods, making a direct comparison of the performance metrics irrational. Even the risk of patient-level overlaps between training and validation or testing sets is notable. Retrospective validation has been adopted in almost all studies; however, validation performed in a prospective manner wins more confidence over the

reliability and applicability of such a proposed system, as in other medical domains. Much of the shortcomings directly or indirectly relate to the scarcity of large open academic datasets incorporating population dynamics and annotated using the most recent standard terminology. An end-to-end holistic solution, covering all possible medical conditions and population diversity with the capability of taking in the raw frames and suggesting disease level diagnosis based on learned clinical relevance established by individual frame-level findings, promises a significant impact, and could be a possible future direction of work for researchers in this domain. Explainability and transparency are also less explored areas in VCE analysis. Existing ethical and legal bindings narrow the scope of possibilities in which AI can potentially leverage healthcare facilities. Regulations regarding ethical and legal concerns need to be updated to provide precise guidance on the extent of explainability required for AI-based diagnostics systems, data privacy, security, and usability issues. Overcoming the contemporary challenges reviewed in this paper promises great potential for VCE to become a first-line gold standard, not just for the small bowel but also for other organs of the GI tract.

CRediT authorship contribution statement

Haroon Wahab: Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Investigation, Formal analysis. **Irfan Mehmood:** Investigation, Data curation, Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Funding acquisition. **Hassan Ugail:** Validation, Formal analysis, Conceptualization, Writing – review & editing, Writing – original draft, Investigation. **Arun Kumar Sangaiah:** Conceptualization, Investigation, Writing – review & editing. **Khan Muhammad:** Conceptual review & editing, Revision review, Writing – review & editing, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research reported here was funded by the Commonwealth Scholarship Commission and the Foreign Commonwealth and Development Office in the UK (grant reference: PKCS-2021-643). The authors are grateful for their support. All views expressed here are those of the authors, not the funding body.

References

[1] A.D. Sperber, et al., Worldwide prevalence and burden of functional gastrointestinal disorders, results of rome foundation global study, Gastroenterology 160 (1) (2021) 99–114.e3, http://dx.doi.org/10.1053/j.gastro.2020.04.014.

[2] Guts Charity, Digesting the facts: What people are thinking about their digestive health, 2016, [Online]. Available: http://gutscharity.org.uk/wp-content/uploads/2016/08/DigestingTheFactsReport.pdf.

[3] M.F. Kaminski, et al., Quality indicators for colonoscopy and the risk of interval cancer, N Engl. J. Med. 362 (19) (2010) 1795–1803, http://dx.doi.org/10.1056/NEJMoa0907667, (in eng).

[4] S.Y. Kim, H.S. Kim, H.J. Park, Adverse events related to colonoscopy: Global trends and future challenges, World J. Gastroenterol. 25 (2) (2019) 190–204, http://dx.doi.org/10.3748/wjg.v25.i2.190, (in eng).

[5] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, Nature 405 (6785) (2000) 417, http://dx.doi.org/10.1038/35013140, (in eng).

[6] M. plc. PILLCAM SB 3 System, 2022, https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-sb-3-system.html, (accessed 2022).

[7] J.E. Axelrad, S. Lichtiger, V. Yajnik, Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment, World J. Gastroenterol. 22 (20) (2016) 4794–4801, http://dx.doi.org/10.3748/wjg.v22.i20.4794, (in eng).

[8] J. Yu, et al., Inflammatory bowel disease and risk of adenocarcinoma and neuroendocrine tumors in the small bowel, Ann. Oncol. 33 (6) (2022) 649–656.

[9] Z. Ding, et al., Gastroenterology 157 (4) (2019) 1044–1054.e5, 2022/01/11.

[10] A. Khan, A. Sohail, U. Zahoora, A.S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, 53, (8) 2020, pp. 5455–5516.

[11] D.H. Hubel, T.N. Wiesel, Receptive fields,binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154, 2022/01/12.

[12] D.H. Hubel, T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, J. Physiol. 195 (1) (1968) 215–243, 2022/01/12.

[13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2020) 4–24.

[14] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[15] S. Soffer, et al., Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis, Gastrointest. Endosc. 92 (4) (2020) 831–839.e8, http://dx.doi.org/10.1016/j.gie.2020.04.039, (in eng).

[16] S.H. Kim, Y.J. Lim, Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges, Diagnostics (Basel) 11 (9) (2021) http://dx.doi.org/10.3390/diagnostics11091722, (in eng).

[17] R. Trasolini, M.F. Byrne, Artificial intelligence and deep learning for small bowel capsule endoscopy, Dig. Endosc. 33 (2) (2021) 290–297, http://dx.doi.org/10.1111/den.13896, (in eng).

[18] Y. Hwang, J. Park, Y.J. Lim, H.J. Chun, Application of artificial intelligence in capsule endoscopy: Where are we now? Clin. Endoscopy 51 (6) (2018) 547–551.

[19] Y.J. Yang, The future of capsule endoscopy: The role of artificial intelligence and other technical advancements, Clin. Endoscopy 53 (4) (2020) 387–394.

[20] X. Dray, E. Toth, T. de Lange, A. Koulaouzidis, Artificial intelligence, capsule endoscopy, databases, and the Sword of Damocles, Endosc. Int. Open 9 (11) (2021) E1754–e1755, http://dx.doi.org/10.1055/a-1521-4882, (in eng).

[21] M. Alagappan, J.R.G. Brown, Y. Mori, T.M. Berzin, Artificial intelligence in gastrointestinal endoscopy: The future is almost here, World J. Gastrointest. Endosc. 10 (10) (2018) 239–249, http://dx.doi.org/10.4253/wjge.v10.i10.239, (in eng).

[22] K. Muhammad, S. Khan, N. Kumar, J. Del Ser, S. Mirjalili, Vision-based personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges, 113, 2020, pp. 266–280.

[23] K. Namikawa, et al., Utilizing artificial intelligence in endoscopy: a clinician's guide, Expert Rev. Gastroenterol. Hepatol. 14 (8) (2020) 689–706, http://dx.doi.org/10.1080/17474124.2020.1779058, (in eng).

[24] A. Adadi, A survey on data-efficient algorithms in big data era, 8, (1) 2021, p. 24.

[25] I. Ogobuiro, G. Justin, T. Faiz, Physiology, Gastrointestinal. [Updated 2021 Apr 25]. In: StatPearls [Internet].

[26] I. Sensoy, A review on the food digestion in the digestive tract and the used in vitro models, 4, 2021, pp. 308–319.

[27] P. Swain, A. Fritscher-Ravens, Role of video endoscopy in managing small bowel disease, Gut 53 (12) (2004) 1866–1875.

[28] D.G. Hewett, C.J. Kahi, D.K. Rex, Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointest. Endosc. Clin. N. Am. 20 (4) (2010) 673–684, http://dx.doi.org/10.1016/j.giec.2010.07.011, (in eng).

[29] A. Koffas, F.-M. Laskaratos, O. Epstein, Training in video capsule endoscopy: Current status and unmet needs, World J. Gastrointest. Endoscopy 11 (6) (2019) 395–402.

[30] A. Sieg, Capsule endoscopy compared with conventional colonoscopy for detection of colorectal neoplasms, World J. Gastrointest. Endoscopy 3 (5) (2011) 81–85.

[31] S.H. Kim, H.J. Chun, Capsule endoscopy: Pitfalls and approaches to overcome, Diagnostics (Basel, Switzerland) 11 (10) (2021) 1765.

[32] Y. Zheng, J. Hawkins L. Fau-Wolff, O. Wolff J. Fau-Goloubeva, E. Goloubeva O. Fau-Goldberg, E. Goldberg, Detection of lesions during capsule endoscopy: physician performance is disappointing, no. 1572-0241 (Electronic).

[33] E. Rondonotti, et al., Can we improve the detection rate and interobserver agreement in capsule endoscopy?, no. 1878-3562 (Electronic).

[34] D.R. Cave, S. Hakimian, K. Patel, Current controversies concerning capsule endoscopy, Dig. Dis. Sci. 64 (11) (2019) 3040–3047, http://dx.doi.org/10.1007/s10620-019-05791-4, (in eng).

[35] L.Y. Korman, et al., Capsule endoscopy structured terminology (CEST): proposal of a standardized and structured terminology for reporting capsule endoscopy procedures, Endoscopy 37 (10) (2005) 951–959, http://dx.doi.org/10.1055/s-2005-870329, (in eng).
[36] D.J. Oh, Y. Hwang, Y.J. Lim, A current and newly proposed artificial intelligence algorithm for reading small bowel capsule endoscopy, Diagnostics (Basel, Switzerland) 11 (7) (2021) 1183.
[37] L.H. Lai, G.L. Wong, D.K. Chow, J.Y. Lau, J.J. Sung, W.K. Leung, Inter-observer variations on interpretation of capsule endoscopies, Eur. J. Gastroenterol. Hepatol. 18 (3) (2006) 283–286, http://dx.doi.org/10.1097/00042737-200603000-00009, (in eng).
[38] B.S. Lewis, The great mimickers, Techniques Gastrointest. Endoscopy 8 (4) (2006) 175–181.
[39] R. Leenhardt, et al., Nomenclature and semantic description of vascular lesions in small bowel capsule endoscopy: an international delphi consensus statement, Endoscopy Int. Open 7 (3) (2019) E372–E379.
[40] R. Leenhardt, et al., Nomenclature and semantic descriptions of ulcerative and inflammatory lesions seen in Crohn's disease in small bowel capsule endoscopy: An international Delphi consensus statement, United Eur. Gastroenterol. J. 8 (1) (2020) 99–107.
[41] R. Leenhardt, et al., A guide for assessing the clinical relevance of findings in small bowel capsule endoscopy: analysis of 8064 answers of international experts to an illustrated script questionnaire, Clin. Res. Hepatol. Gastroenterol. 45 (6) (2021) 101637, http://dx.doi.org/10.1016/j.clinre.2021.101637, (in eng).
[42] M. Marquès Camí, A. Serracarbasa, G. D'Haens, M. Löwenberg, Characterization of mucosal lesions in crohn's disease scored with capsule endoscopy: A systematic review, Front. Med. Systematic Rev. 7 (2021) http://dx.doi.org/10.3389/fmed.2020.600095, (in English).
[43] C. Spada, M.-E. Riccioni, R. Urgesi, G. Costamagna, Capsule endoscopy in celiac disease, World J. Gastroenterol. 14 (26) (2008) 4146–4151.
[44] F. Branchi, et al., Small-bowel capsule endoscopy in patients with celiac disease, axial versus lateral/panoramic view: Results from a prospective randomized trial, Dig. Endosc. 32 (5) (2020) 778–784, http://dx.doi.org/10.1111/den.13575, (in eng).
[45] M.S. Chang, M. Rubin, S.K. Lewis, P.H. Green, Diagnosing celiac disease by video capsule endoscopy (VCE) when esophogastroduodenoscopy (EGD) and biopsy is unable to provide a diagnosis: a case series, 12, (1) 2012, p. 90.
[46] D.K. Christodoulou, D.E. Sigounas, K.H. Katsanos, G. Dimos, E.V. Tsianos, Small bowel parasitosis as cause of obscure gastrointestinal bleeding diagnosed by capsule endoscopy, World J. Gastrointest. Endoscopy 2 (11) (2010) 369–371.
[47] E. Akin, O. Ersoy, Capsule endoscopy in celiac disease, in: Gastroenterology Research and Practice, Vol. 2012, 2012, p. 676073.
[48] B. Rosa, et al., Scoring systems in clinical small-bowel capsule endoscopy: all you need to know!, Endosc. Int. Open 9 (6) (2021) E802–e823, http://dx.doi.org/10.1055/a-1372-4051, (in eng).
[49] P.H. Smedsrud, et al., Kvasir-capsule, a video capsule endoscopy dataset, 8, (1) 2021, p. 142.
[50] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[51] A. Nemeth, G. Wurm Johansson, J. Nielsen, H. Thorlacius, E. Toth, Capsule retention related to small bowel capsule endoscopy: a large European single-center 10-year clinical experience, United Eur. Gastroenterol. J. 5 (5) (2016) 677–686, 2022/10/29.
[52] A. Koulaouzidis, et al., KID project: an internet-based digital video atlas of capsule endoscopy for research purposes, Endosc. Int. Open 5 (6) (2017) E477–e483, http://dx.doi.org/10.1055/s-0043-105488, (in eng).
[53] J. Bernal, H. Aymeric, Gastrointestinal Image ANAlysis (GIANA) Angiodysplasia D & L challenge.
[54] R. Leenhardt, et al., CAD-CAP: a 25, 000-image database serving the development of artificial intelligence for capsule endoscopy, Endosc. Int. Open 8 (3) (2020) E415–E420.
[55] Gastrolab, The gastrointestinal site, 2022, http://www.gastrolab.net/index.htm, (accessed 2022).
[56] M. Keuchel, S.O. Al-Harthi, F. Hagenmueler, New automatic mode of rapid 4 software reduces reading time for small bowel pillcam studies, in: Intl. Conf. on Capsule Endoscope, 2006.
[57] J. Westerhof, J.J. Koornstra, R.K. Weersma, Can we reduce capsule endoscopy reading times?, 69, (3) 2009, pp. 497–502, Part 1.
[58] H. Liu, N. Pan, H. Lu, E. Song, Q. Wang, C.-C. Hung, Wireless capsule endoscopy video reduction based on camera motion estimation, J. Digital Imaging 26 (2) (2013) 287–301.
[59] D.K. Iakovidis, S. Tsevas, A. Polydorou, Reduction of capsule endoscopy reading times by unsupervised image mining, Comput. Med. Imaging Graph. 34 (6) (2010) 471–478, http://dx.doi.org/10.1016/j.compmedimag.2009.11.005, (in eng).
[60] H.-G. Lee, M.-K. Choi, S.-C. Lee, Motion analysis for duplicate frame removal in wireless capsule endoscope, in: Medical Imaging 2011: Image Processing, Vol. 7962, International Society for Optics and Photonics, 2011, 79621T.
[61] H.G. Lee, M.K. Choi, B.S. Shin, S.C. Lee, Reducing redundancy in wireless capsule endoscopy videos, Comput. Biol. Med. 43 (6) (2013) 670–682, http://dx.doi.org/10.1016/j.compbiomed.2013.02.009, (in eng).
[62] M. Drozdzal, L. Igual, J. Vitrià, C. Malagelada, F. Azpiroz, P. Radeva, Aligning endoluminal scene sequences in wireless capsule endoscopy, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 117–124.
[63] I. Mehmood, M. Sajjad, S.W. Baik, Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure, J. Med. Syst. 38 (9) (2014) 109, http://dx.doi.org/10.1007/s10916-014-0109-y, (in eng).
[64] I. Mehmood, M. Sajjad, S.W. Baik, Mobile-cloud assisted video summarization framework for efficient management of remote sensing data generated by wireless capsule sensors, Sensors 14 (9) (2014).
[65] N. Ejaz, I. Mehmood, S.W. Baik, MRT letter: visual attention driven framework for hysteroscopy video abstraction, Microsc. Res. Tech. 76 (6) (2013) 559–563, http://dx.doi.org/10.1002/jemt.22205, (in eng).
[66] X. Chu, et al., Epitomized summarization of wireless capsule endoscopic videos for efficient visualization, in: presented at the Proceedings of the 13th international conference on Medical image computing and computer-assisted intervention: Part II, Beijing, China, 2010.
[67] J. Chen, Y. Zou, Y. Wang, Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1303–1308.
[68] L. Lan, C. Ye, Recurrent generative adversarial networks for unsupervised WCE video summarization, 222, 2021, 106971.
[69] B. Sushma, P. Aparna, Summarization of wireless capsule endoscopy video using deep feature matching and motion analysis, IEEE Access 9 (2021) 13691–13703.
[70] F. Phillips, S. Beg, Video capsule endoscopy: pushing the boundaries with software technology, Translational Gastroenterol. Hepatol. 6 (2020).
[71] D.K. Iakovidis, S.V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, V.P. Plagianakos, Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification, IEEE Trans. Med. Imaging 37 (10) (2018) 2196–2210.
[72] S. Fan, L. Xu, Y. Fan, K. Wei, L. Li, Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images, Phys. Med. Biol. 63 (16) (2018) 165001, http://dx.doi.org/10.1088/1361-6560/aad51c, (in eng).
[73] J.Y. He, X. Wu, Y.G. Jiang, Q. Peng, R. Jain, Hookworm detection in wireless capsule endoscopy images with deep learning, IEEE Trans. Image Process. 27 (5) (2018) 2379–2392, http://dx.doi.org/10.1109/tip.2018.2801119, (in eng).
[74] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, U.J. Tanik, Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images, J. Exp. Theoret. Artif. Intell. 33 (4) (2021) 577–599.
[75] R. Leenhardt, et al., A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy, Gastrointest. Endosc. 89 (1) (2019) 189–194, http://dx.doi.org/10.1016/j.gie.2018.06.036, (in eng).
[76] T. Aoki, et al., Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network, Gastrointest. Endosc. 89 (2) (2019) 357–363.e2, http://dx.doi.org/10.1016/j.gie.2018.10.027, (in eng).
[77] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, D. Al-Jumeily, Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images, Sensors (Basel) 19 (6) (2019) http://dx.doi.org/10.3390/s19061265, (in eng).
[78] P.M. Vieira, N.R. Freitas, J. Valente, I.F. Vaz, C. Rolanda, C.S. Lima, Automatic detection of small bowel tumors in wireless capsule endoscopy images using ensemble learning, Med. Phys. 47 (1) (2020) 52–63, http://dx.doi.org/10.1002/mp.13709, (in eng).
[79] G. Pan, G. Yan, X. Qiu, J. Cui, Bleeding detection in wireless capsule endoscopy based on probabilistic neural network, J. Med. Syst. 35 (6) (2011) 1477–1484, http://dx.doi.org/10.1007/s10916-009-9424-0, (in eng).
[80] J. Xiao, M.Q. Meng, A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (2016) 639–642, http://dx.doi.org/10.1109/embc.2016.7590783, (in eng).
[81] Y. Yuan, B. Li, M.Q. Meng, Bleeding frame and region detection in the wireless capsule endoscopy video, IEEE J. Biomed. Health Inform. 20 (2) (2016) 624–630, http://dx.doi.org/10.1109/jbhi.2015.2399502, (in eng).

[82] P. Li, Z. Li, F. Gao, L. Wan, J. Yu, Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 1518–1523.
[83] X. Jia, M.Q. Meng, A study on automated segmentation of blood regions in Wireless Capsule Endoscopy images using fully convolutional networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 179–182.
[84] T. Aoki, et al., Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network, J. Gastroenterol. Hepatol. 35 (7) (2020) 1196–1200, http://dx.doi.org/10.1111/jgh.14941, (in eng).
[85] M. Mascarenhas Saraiva, et al., Artificial Intelligence and Capsule Endoscopy: Automatic Detection of Small Bowel Blood Content using a Convolutional Neural Network, 2021.
[86] F. Rustam, et al., Wireless capsule endoscopy bleeding images classification using CNN based model, IEEE Access 9 (2021) 33675–33688.
[87] P. Laiz, J. Vitrià, H. Wenzek, C. Malagelada, F. Azpiroz, S. Seguí, WCE polyp detection with triplet based embeddings, 86, 2020, 101794.
[88] Y. Yuan, M.Q. Meng, Deep learning for polyp recognition in wireless capsule endoscopy images, Med. Phys. 44 (4) (2017) 1379–1389, http://dx.doi.org/10.1002/mp.12147, (in eng).
[89] A. Karargyris, N. Bourbakis, Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos, IEEE Trans. Biomed. Eng. 58 (10) (2011) 2777–2786, http://dx.doi.org/10.1109/tbme.2011.2155064, (in eng).
[90] B. Li, M.Q. Meng, Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection, IEEE Trans. Inf. Technol. Biomed. 16 (3) (2012) 323–329.
[91] E.S. Nadimi, et al., Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy, 81, 2020, 106531.
[92] D. Jha, et al., NanoNet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), 2021, pp. 37–43.
[93] Y. Yuan, J. Wang, B. Li, M.Q.H. Meng, Saliency based ulcer detection for wireless capsule endoscopy diagnosis, IEEE Trans. Med. Imaging 34 (10) (2015) 2046–2057.
[94] S. Wang, Y. Xing, L. Zhang, H. Gao, H. Zhang, A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks, Phys. Med. Biol. 64 (23) (2019) 235014, http://dx.doi.org/10.1088/1361-6560/ab5086, (in eng).
[95] A. Tsuboi, et al., Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images, Dig. Endosc. 32 (3) (2020) 382–390, http://dx.doi.org/10.1111/den.13507, (in eng).
[96] T. Zhou, et al., Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method, Comput. Biol. Med. 85 (2017) 1–6, http://dx.doi.org/10.1016/j.compbiomed.2017.03.031, (in eng).
[97] M.A. Khan, et al., Computer-aided gastrointestinal diseases analysis from wireless capsule endoscopy: A framework of best features selection, IEEE Access 8 (2020) 132850–132859.
[98] J. Xia, et al., Use of artificial intelligence for detection of gastric lesions by magnetically controlled capsule endoscopy, Gastrointest. Endosc. 93 (1) (2021) 133–139.e4, http://dx.doi.org/10.1016/j.gie.2020.05.027, (in eng).
[99] J. Park, et al., Artificial intelligence that determines the clinical significance of capsule endoscopy images can increase the efficiency of reading, PLoS One 15 (10) (2020) e0241474, http://dx.doi.org/10.1371/journal.pone.0241474, (in eng).
[100] E. Klang, et al., Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy, Gastrointest. Endosc. 91 (3) (2020) 606–613.e2, http://dx.doi.org/10.1016/j.gie.2019.11.012, (in eng).
[101] H. Saito, et al., Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network, Gastrointest. Endosc. 92 (1) (2020) 144–151.e1, http://dx.doi.org/10.1016/j.gie.2020.01.054, (in eng).
[102] M.J. Mascarenhas Saraiva, et al., Deep learning and capsule endoscopy: automatic identification and differentiation of small bowel lesions with distinct haemorrhagic potential using a convolutional neural network, BMJ Open Gastroenterol. 8 (1) (2021) http://dx.doi.org/10.1136/bmjgast-2021-000753, (in eng).
[103] Z. Xiao, L.N. Feng, A study on wireless capsule endoscopy for small intestinal lesions detection based on deep learning target detection, IEEE Access 8 (2020) 159017–159026.
[104] Y. Hwang, et al., Improved classification and localization approach to small bowel capsule endoscopy using convolutional neural network, Digestive Endosc. 33 (4) (2021) 598–607, 2022/02/22.
[105] S. Adewole, et al., Lesion2Vec: Deep Metric Learning for Few Shot Multiple Lesions Recognition in Wireless Capsule Endoscopy Video, 2021.
[106] P. Muruganantham, S.M. Balakrishnan, Attention Aware Deep Learning Model for Wireless Capsule Endoscopy Lesion Classification and Localization, 2022.
[107] A. Vats, M. Pedersen, A. Mohammed, Ø. Hovde, Learning more for Free - a Multi Task Learning Approach for Improved Pathology Classification in Capsule Endoscopy, 2021, pp. 3–13.
[108] Z. Ding, et al., Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model, Gastroenterology 157 (4) (2019) 1044–1054.e5, http://dx.doi.org/10.1053/j.gastro.2019.06.025, (in eng).
[109] T. Aoki, et al., Automatic detection of various abnormalities in capsule endoscopy videos by a deep learning-based system: a multicenter study, Gastrointest. Endosc. 93 (1) (2021) 165–173.e1, 2022/02/22.
[110] S.H. Kim, et al., Efficacy of a comprehensive binary classification model using a deep convolutional neural network for wireless capsule endoscopy, 11, (1) 2021, 17479.
[111] Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, W.J. Deng, Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network, in: 2015 IEEE International Conference on Digital Signal Processing (DSP), 2015, pp. 1274–1278.
[112] S. Seguí, et al., Generic feature learning for wireless capsule endoscopy analysis, 79, 2016, pp. 163–172.
[113] Z. Ding, et al., Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model, Gastroenterology 157 (4) (2019) 1044–1054.e5, 2022/01/11.
[114] M. Flückiger, B.J. Nelson, Ultrasound emitter localization in heterogeneous media, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2007 (2007) 2867–2870, http://dx.doi.org/10.1109/iembs.2007.4352927, (in eng).
[115] F. Carpi, N. Kastelein, M. Talcott, C. Pappone, Magnetically controllable gastrointestinal steering of video capsules, IEEE Trans. Biomed. Eng. 58 (2) (2011) 231–234, http://dx.doi.org/10.1109/tbme.2010.2087332, (in eng).
[116] H. Keller, et al., Method for navigation and control of a magnetically guided capsule endoscope in the human stomach, in: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), 2012, pp. 859–865.
[117] S. Yim, M. Sitti, 3-D localization method for a magnetically actuated soft capsule endoscope and its applications, IEEE Trans. Robot. 29 (5) (2013) 1139–1151, http://dx.doi.org/10.1109/tro.2013.2266754, (in eng).
[118] D. Son, S. Yim, M. Sitti, A 5-D localization method for a magnetically manipulated untethered robot using a 2-D array of hall-effect sensors, IEEE/ASME Trans. Mechatronics 21 (2) (2016) 708–716.
[119] M. Turan, Y. Almalioglu, H. Araújo, E. Konukoglu, M. Sitti, Deep endovo: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots, Neurocomputing 275 (2018) 1861–1870.
[120] C. Jakob Engel, Thomas Schps, Daniel, LSD-SLAM: Large-Scale Direct Monocular SLAM, 2014.
[121] R. Mur-Artal, J.M.M. Montiel, J.D. Tardós, ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.
[122] G. Bao, K. Pahlavan, L. Mi, Hybrid localization of microrobotic endoscopic capsule inside small intestine by data fusion of vision and RF sensors, IEEE Sens. J. 15 (5) (2015) 2669–2678.
[123] Medtronic, PILLCAM™ SB 3 SYSTEM, 2022, https://www.medtronic.com/covidien/en-us/products/capsule-endoscopy/pillcam-sb-3-system.html, (accessed 2022).
[124] Olympus, Endocapsule 10 system, 2022, https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html, (accessed 2022).
[125] S. Beg, et al., Use of rapid reading software to reduce capsule endoscopy reading times while maintaining accuracy, Gastrointest. Endosc. 91 (6) (2020) 1322–1327, http://dx.doi.org/10.1016/j.gie.2020.01.026, (in eng).
[126] I. Goodfellow, et al., Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[127] A. Aggarwal, M. Mittal, G. Battineni, Generative adversarial network: An overview of theory and applications, 1, (1) 2021, 100004.
[128] X. Li, Y. Jiang, J.J. Rodriguez-Andina, H. Luo, S. Yin, O. Kaynak, When medical images meet generative adversarial network: recent development and research opportunities, 1, (1) 2021, p. 5.
[129] A. DuMont Schütte, et al., Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation, 4, (1) 2021, p. 141.
[130] R.J. Chen, M.Y. Lu, T.Y. Chen, D.F.K. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, 5, (6) 2021, pp. 493–497.
[131] A. Tucker, Z. Wang, Y. Rotalinti, P. Myles, Generating high-fidelity synthetic patient data for assessing machine learning healthcare software, 3, (1) 2020, p. 147.

[132] K. Koshino, et al., Narrative review of generative adversarial networks in medical and molecular imaging, Ann. Transl. Med. 9 (9) (2021) 821, http://dx.doi.org/10.21037/atm-20-6325, (in eng).
[133] A. Arora, A. Arora, Synthetic patient data in health care: a widening legal loophole, Lancet 399 (10335) (2022) 1601–1602, 2022/10/29.
[134] S.R. Islam, W. Eberle, S.K. Ghafoor, Towards quantification of explainability in explainable artificial intelligence methods, 2020, arXiv, abs/1911.10104.
[135] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, 10, (1) 2019, p. 1096.
[136] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, PLOS Med. 15 (11) (2018) e1002683.
[137] A.B. Arrieta, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[138] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[139] A. Rai, Explainable AI: From black box to glass box, J. Acad. Mark. Sci. 48 (1) (2020) 137–141.
[140] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer Nature, 2019.
[141] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, 77, 2022, pp. 29–52.
[142] S.G. Kim, N. Theera-Ampornpunt, C.-H. Fang, M. Harwani, A. Grama, S. Chaterji, Opening up the blackbox: an interpretable deep neural network-based classifier for cell-type specific enhancer predictions, BMC Syst. Biol. 10 (2) (2016) 243–258.
[143] G. Yang, F. Raschke, T.R. Barrick, F.A. Howe, Manifold learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering, Magn. Reson. Med. 74 (3) (2015) 868–878.
[144] L.P. Zhao, H. Bolouri, Object-oriented regression for building predictive models with high dimensional omics data from translational studies, J. Biomed. Inform. 60 (2016) 431–445.
[145] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.
[146] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating Shapley values of local components, in: Explainable AI in Healthcare and Medicine, Springer, 2021, pp. 261–270.
[147] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
[148] H. Lee, et al., An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets, Nat. Biomed. Eng. 3 (3) (2019) 173–182.
[149] P. Rajpurkar, et al., AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining, Sci. Rep. 10 (1) (2020) 1–7.
[150] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[151] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[152] M.S. Kovalev, L.V. Utkin, E.M. Kasimov, Survlime: A method for explaining machine learning survival models, Knowl.-Based Syst. 203 (2020) 106164.
[153] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, M.-A. Sicilia, Traceability for trustworthy AI: A review of models and tools, Big Data Cogn. Comput. 5 (2) (2021).
[154] H. Felzmann, E. Fosch-Villaronga, C. Lutz, A. Tamò-Larrieux, Towards transparency by design for artificial intelligence, 26, (6) 2020, pp. 3333–3361.
[155] C. Hashemi-Pour, Interpretability-and-Explainability-Can-Lead-to-more-Reliable-ML, TechTarget, AI Technologies, 2022, Available https://www.techtarget.com/searchenterpriseai/feature/Interpretability-and-explainability-can-lead-to-more-reliable-ML.
[156] S. Gerke, T. Minssen, G. Cohen, Ethical and legal challenges of artificial intelligence-driven healthcare, Artif. Intell. Healthcare (2020) 295–336.
[157] C. Panigutti, A. Monreale, G. Comandè, D. Pedreschi, Ethical, societal and legal issues in deep learning for healthcare, in: Deep Learning in Biology and Medicine, World Scientific (Europe), 2021, pp. 265–313.
[158] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, Q. c. the Precise, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, 20, (1) 2020, p. 310.
[159] B. Murdoch, Privacy and artificial intelligence: challenges for protecting health information in a new era, 22, (1) 2021, p. 122.
[160] S. Carter, K. Win, L. Wang, W. Rogers, B. Richards, N. Houssami, 65 Ethical, legal and social implications of artificial intelligence systems for screening and diagnosis, BMJ Evidence-Based Med. 24 (Suppl 2) (2019) A37.
[161] M. Beil, I. Proft, D. van Heerden, S. Sviri, P.V. van Heerden, Ethical considerations about artificial intelligence for prognostication in intensive care, 7, (1) 2019, p. 70.
[162] D. Cossins, Discriminating algorithms: 5 times AI showed prejudice, 2022, NewScientist. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice (accessed 30 October, 2022).
[163] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (6464) (2019) 447–453.
[164] C. Ross, I. Swetlitz, IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show, 2022, STAT. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments, (accessed 30 October, 2022).
[165] U.H. Security, US Department of Homeland Security, Cybersecurity, 2022, https://www.dhs.gov/topic/cybersecurity, (accessed 30 October, 2022).
[166] U. F. D. A. (FDA), Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper and Request for Feedback, Ed, 2020.

Haroon Wahab is a Commonwealth Scholar (funded by UK govt.), and a PhD student at Centre for Visual Computing (CVC), University of Bradford, UK. His areas of research interests are in visual computing, machine learning with applications in healthcare diagnostics and decision support systems.

Irfan Mehmood (SM'16) has been involved in IT industry and academia in Pakistan, South Korea, and UK for over 10 years. He is now serving as a Assistant Professor of Applied Artificial Intelligence, School of Electrical Engineering and Computer Science, Media, Design and Technology, University of Bradford, UK. His sustained contribution at various research and industry-collaborative projects gives him an extra edge to meet the current challenges faced in the field of multimedia analytics. Specifically, he has made significant contribution in the areas of visual surveillance, information mining and data encryption. He has published 90+ papers in peer-reviewed international journals and conferences such as Information Fusion, Neurocomputing, IEEE Access, IEEE Transactions on Industrial Informatics, IEEE Internet of Things Journal, International Journal of Information Management, Future Generation Computer Systems, Sensors, Journal of Visual Communication and Image Representation, Multimedia Tools and Applications, Computers in Human Behavior, EURASIP Journal on Image and Video Processing, Mobile Networks and Applications, Computers in Biology and Medicine, Journal of Medical Systems, Signal, Image and Video Processing, Bio-Medical Materials and Engineering, KSII Transactions on Internet and Information Systems, NBIS 2015, MITA 2015, PlatCon 2016, SKIMA 2019, and IWFCV 2020. He is serving as a professional reviewer for numerous well-reputed journals such as Journal of Visual Communication and Image Representation, Future Generation Computer Systems, IEEE Access, Journal of SuperComputing, Signal Image and Video Processing, Multimedia Tools and Applications, ACM Transactions on Embedded Computing Systems, and Enterprise Information Systems. He acted as GE/LGE in several special issues of SCI/SCIE indexed journals and is currently involved in editing of several other special issues. Contact at i.mehmood4@bradford.ac.uk.

Hassan Ugail is the director of the Centre for Visual Computing in the Faculty of Engineering and Informatics at the University of Bradford, UK. He has a first class BSc Honours degree in Mathematics from King's College London and a PhD in the field of geometric design from the School of Mathematics at the University of Leeds. Professor Ugail's research interests include computer based geometric and functional design, imaging and machine learning

Arun Kumar Sangaiah received his Ph.D. from School of Computer Science and Engineering, VIT University, Vellore, India. He is currently a Full Professor with National Yunlin University of Science and Technology, Taiwan. He has published more than 300 research articles in refereed journals (IEEE TII, IEEE TITS, IEEE TNSE, IEEE TETCI, IEEE SysJ, IEEE SensJ, IEEE IOTJ, ACM TOSN) 11 edited books, as well as 1 patents (held and filed) and 3 projects, among one funded by Ministry of IT of India and few international projects (CAS, Guangdong Research fund, Australian Research Council) cost worth of 500000 USD. Dr. Sangaiah has received many awards, Clarivate Highly Cited Researcher, Yushan Young Scholar fellowship, Top 2% Scientist, PIFI-CAS fellowship, Top-10 outstanding researcher, CSI significant Contributor etc. Also, he is responsible for Editor-in-Chief, and Associate Editor of various reputed ISI journals. Dr. Sangaiah is a visiting scientist (2018-2019) with Chinese Academy of Sciences (CAS), China and visiting researcher of Université Paris-Est (UPEC), France (2019-2020) and etc.

Khan Muhammad [S'16, M'18, SM'22] received his Ph.D. degree in Digital Contents from Sejong University, South Korea in 2019. He is currently the director of Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab) and an Assistant Professor with the Department of Applied AI, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul, South Korea. His research interests include intelligent video surveillance, medical image analysis, information security, video summarization, multimedia data analysis, computer vision, IoT/IoMT, and smart cities. He has registered 10 patents and has contributed 220+ papers in peer-reviewed journals and conference proceedings in his areas of research. He is an Associate Editor/Editorial Board Member of more than 15 journals. He is among the highly cited researchers in 2021 and 2022 according to the Web of Science (Clarivate).