PolypInsight: A Unique Unsupervised Deep Learning Framework for Cancer Detection in Colonoscopy Without Human Labels

1st Dr. Satyabrata Roy
dept. of CSE
Manipal University Jaipur
Jaipur, India
satyabrata.roy@jaipur.manipal.edu

2nd Akash Raj
dept. of CSE
Manipal University Jaipur
Jaipur, India
akash.229301533@muj.manipal.edu

3rd Sainyam Acharya
dept. of CSE
Manipal University Jaipur
Jaipur, India
sainyam.229301524@muj.manipal.edu

***Abstract*—Colorectal cancer is a major global health concern, and early identification of precancerous colon polyps is crucial for improving patient prognosis. In this paper, we propose a unique and previously unexplored hybrid framework that integrates state-of-the-art deep learning-based detection with unsupervised clustering to classify colon polyps into cancerous and non-cancerous categories—without relying on any explicit labels. We utilize the YOLOv8 object detector to automatically identify and crop polyp regions from colonoscopy video frames in the REAL-Colon dataset. Deep feature representations of the cropped polyps are extracted using the pretrained MobileNetV2 convolutional neural network. To reduce redundancy and highlight informative patterns in the high-dimensional feature space, we apply Variance Thresholding followed by dimensionality reduction techniques—Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). These compact embeddings are then clustered using K-Means and Agglomerative Hierarchical Clustering. We further introduce a novel ensemble clustering mechanism that combines the outputs of both algorithms to enhance cluster reliability. The clustering process aims to uncover meaningful separations between polyp types, such as malignant and benign cases, entirely through unsupervised learning. Internal validation metrics—Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index—are used to assess cluster quality. Visualization through 2D UMAP plots confirms the emergence of distinguishable clusters. Experimental results validate that our pipeline can successfully identify clinically relevant groupings, with ensemble clustering showing superior performance. The proposed methodology represents a promising direction for computer-aided diagnosis in pathology-scarce scenarios and sets the foundation for future enhancements, including spatial feature refinement via SAM and HSV-based filtering.**

***Index Terms*—Colorectal cancer, Polyp detection, Unsupervised clustering, Deep feature extraction, YOLOv8, Dimensionality reduction**

I. Introduction

Colorectal cancer (CRC) is one of the most prevalent malignancies and a leading cause of cancer-related deaths globally [1]. Most colorectal cancers develop from adenomatous polyps in the colon; detecting and removing these polyps during colonoscopy can prevent the progression to cancer. As such, high-quality colonoscopic examination and accurate polyp diagnosis are critical for early intervention [2]. However, distinguishing between polyps that are benign (non-neoplastic) and those with malignant potential (adenomas that could progress to cancer) can be challenging during the procedure. Traditionally, polyp classification relies on histopathological analysis after polyp removal, but real-time assessment during endoscopy (optical diagnosis) is a desirable goal to inform immediate clinical decisions.

In recent years, artificial intelligence (AI) and deep learning have shown great promise in assisting gastroenterologists by automatically detecting and characterizing polyps from colonoscopy imagery [3]. Convolutional neural networks (CNNs) trained on large endoscopic datasets can achieve high sensitivity in polyp detection, serving as a second observer to alert endoscopists to polyps that might be missed. Beyond detection (often referred to as computer-aided detection, CADe), there is growing interest in computer-aided diagnosis (CADx), where AI models predict polyp histology (e.g., adenomatous vs hyperplastic) based on visual features [4]. Most CADx approaches to date have been supervised, requiring extensive labeled data of polyps with known outcomes. Obtaining such annotations is labor-intensive and requires expert pathology correlation.

In this paper, we propose a novel hybrid framework that combines state-of-the-art deep learning for polyp detection with unsupervised clustering and ensemble clustering techniques for polyp classification. Our approach is designed to leverage the strengths of CNN-based feature extraction while addressing scenarios where explicit labels (e.g., histopathology results) may be limited or unavailable for training a classifier. We first detect polyps in colonoscopy videos using the YOLOv8 object detector, which provides real-time performance and high accuracy for object localization. Detected polyp regions, labeled initially as polyp and non-polyp, are then cropped, and only the identified polyp regions are retained for further analysis. These cropped polyp images are passed through pretrained CNN feature extractors (MobileNetV2) to obtain rich feature vectors encoding the visual characteristics of the polyps. Instead of directly applying a supervised clas-

sifier, we employ unsupervised clustering algorithms to partition these feature vectors into groups, specifically aiming to differentiate potentially cancerous from non-cancerous polyps without explicit supervision.

We utilize K-Means and Agglomerative Hierarchical clustering techniques separately to identify inherent groupings in the data. Subsequently, we apply ensemble clustering, which combines the results of K-Means and Agglomerative methods to produce a robust and consensus-based partitioning of the data. The hypothesis is that integrating multiple clustering outcomes can improve the reliability and interpretability of the classification results, better corresponding to meaningful categories of polyps (such as neoplastic vs. non-neoplastic).

Our contributions are as follows: (1) We introduce a fully automated pipeline for colon polyp analysis that integrates detection and ensemble clustering, enabling exploration of polyp feature space structure on the large-scale REAL-Colon dataset [5]. (2) We extract deep features exclusively from the MobileNetV2 CNN architecture to effectively represent polyp appearance. (3) We implement and evaluate multiple clustering techniques (K-Means, Agglomerative) individually and then enhance them with ensemble clustering, in conjunction with dimensionality reduction (PCA and UMAP), to investigate their capability to differentiate polyps by underlying pathology in an unsupervised manner. (4) We evaluate the quality of the clustering results using internal validation metrics (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) and provide visual interpretations via 2D embedding plots (UMAP). The results and analysis demonstrate that the ensemble clustering approach achieves a notable separation of polyp types, highlighting the potential of unsupervised and ensemble learning as tools for exploratory analysis and aiding medical decision support. Additionally, we discuss the limitations of the current approach and propose directions for future work in developing more robust and clinically useful AI systems for colonoscopy.

II. Related Work

Deep learning has revolutionized medical image analysis in the last decade, and colonoscopy image analysis is no exception. In the area of *polyp detection*, early works relied on handcrafted features and classical computer vision techniques to identify polyp candidates in frames [6]. These included shape-based methods (e.g., detecting elliptical shapes or curvatures characteristic of polyps) and texture or color-based filters. With the advent of CNNs, data-driven approaches have become dominant. Models originally developed for generic object detection, such as Faster R-CNN and the YOLO (You Only Look Once) family, have been applied to colonoscopy videos with considerable success [3]. The YOLO series in particular offers real-time detection capability, which is crucial for live video analysis during endoscopy. YOLOv8, the latest iteration of the YOLO framework at the time of this study, introduces architectural improvements such as an anchor-free detection head, CSP (Cross Stage Partial) backbone enhancements, and mosaic data augmentation, leading to improved accuracy and speed [7]. These properties make YOLOv8 well-suited for the polyp detection task, where both high sensitivity and low latency are desired.

For *polyp classification (CADx)*, research has explored both image-based classification and sequence analysis. Some works use high-definition white-light endoscopy images or narrow-band imaging (NBI) frames and train classifiers (often CNNs) to distinguish adenomas from hyperplastic polyps [4]. Others employ temporal models or ensembles that take advantage of multiple frames or video segments of a polyp. The common thread is supervised learning: methods like ResNet, EfficientNet, or specialized CNN architectures have been trained on labeled datasets where each polyp image or region is annotated with a class label (benign vs malignant). While these approaches have achieved promising accuracy (some reporting above 90% classification rates on benchmark datasets), they heavily rely on the availability of expertly labeled training data. To alleviate labeling costs, there has been interest in weakly supervised and semi-supervised techniques. For instance, some studies use self-supervised learning to pretrain on unlabeled endoscopy images and then fine-tune on a smaller labeled set.

In contrast, purely *unsupervised clustering* for polyp classification has not been extensively studied, to our knowledge. Unsupervised learning has found use in other medical image contexts, such as clustering patient images by disease subtype or discovering novel groupings in imaging genetics, but applying it to endoscopy polyp images is challenging due to subtle visual differences and noise. Nevertheless, clustering techniques like K-Means and Agglomerative clustering have been used as analysis tools to visualize high-dimensional feature spaces and sometimes to generate pseudo-labels for further model training [8]. Visualization methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have been particularly useful for interpreting CNN feature embeddings by projecting them into 2D/3D space where clusters or continuous gradients in the data can be observed. Recent large-scale datasets such as REAL-Colon [5] provide an opportunity to explore unsupervised patterns because they contain a diverse array of polyps with associated metadata (e.g., histology). By leveraging such rich data and powerful feature extractors, it becomes feasible to examine whether an unsupervised approach could separate polyps into meaningful categories without direct supervision.

Table I compares our approach with several related works:

TABLE I
COMPARISON WITH EXISTING WORKS

Feature	Our Approach	Other Works
Detection Method	YOLOv8 (Real-time Object Detection)	YOLOv3 [7], Faster R-CNN [3]
Feature Extraction	MobileNetV2 (Pretrained)	CNNs (ResNet) [9]
Clustering Methods	K-Means, Agglomerative, Ensemble Clustering	K-Means [8], Agglomerative [9]
Dimensionality Reduction	PCA, UMAP for Visualization	t-SNE [8]
Additional Features	No color-based features, only CNN-based	Hybrid Methods [6], Color-based Features [8]
Evaluation Metrics	Silhouette Score, DB Index, CH Index	DB Index, Silhouette Score [8]

Our work builds on these developments. We use the estab-

lished CNN architecture MobileNetV2 [9] to extract features, rather than training a new network from scratch, aligning with practices in transfer learning for medical imaging to handle limited data per class. We incorporate multiple clustering algorithms, each with different assumptions about cluster shape and density, to thoroughly evaluate what structure exists in the polyp feature data. Furthermore, we implement a novel ensemble clustering approach that combines the strengths of K-Means and Agglomerative clustering methods. This ensemble strategy aims to leverage the complementary information provided by different clustering solutions, potentially improving cluster stability and clinical interpretability. Additionally, our framework uniquely focuses solely on CNN-based features extracted from MobileNetV2, without additional color-based features as found in some earlier works. This simplification avoids additional computational complexity and focuses purely on leveraging deep learning-based feature extraction for effective clustering.

By combining these approaches, our framework shifts the paradigm from fully supervised classification to exploratory unsupervised and ensemble clustering, aiming to uncover latent groupings that correspond to clinically relevant categories.

III. Methodology

A. Dataset: REAL-Colon

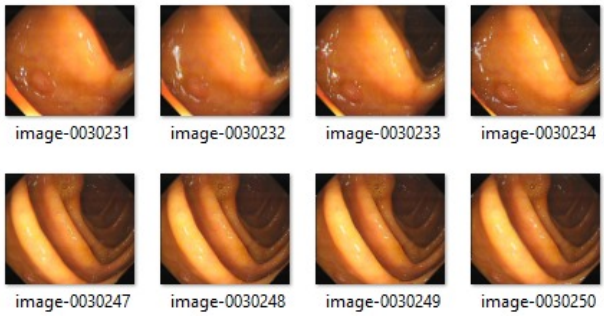

Fig. 1. RealColon dataset sample

We conducted our study using the recently released *REAL-Colon* dataset [5], a large-scale collection of colonoscopy videos intended for developing robust AI models. The REAL-Colon dataset comprises 60 full-length colonoscopy recordings collected from multiple medical centers, totaling approximately 2.7 million video frames at full resolution. Expert gastroenterologists have annotated over 350,000 frames in these videos with bounding boxes around polyps, yielding a comprehensive set of polyp instances under varied imaging conditions (different patients, endoscope devices, bowel preparation quality, etc.). Each detected polyp in the dataset is also associated with clinical metadata, including patient information, polyp size, and importantly, the histopathology result of that polyp (e.g., hyperplastic, adenoma, serrated adenoma, carcinoma, etc.), obtained after its removal and biopsy. This wealth of data makes REAL-Colon an excellent resource for our purposes, as it contains both the visual information and the ground truth labels of polyps, albeit the labels are not used in training any classifier in our unsupervised approach.

For our experiments, we sampled a subset of the REAL-Colon data focusing on distinct polyp instances. We extracted frames where polyps are clearly visible and used the provided bounding box annotations to crop the polyp regions from those frames. In total, we gathered N polyp images (regions of interest) for feature extraction and clustering. This included a mix of polyps with various sizes and types, roughly balanced between those that were ultimately diagnosed as adenomatous (pre-cancerous or cancerous potential) and those that were non-neoplastic (benign). All images were resized to a fixed resolution (e.g., 224×224 pixels) to serve as input to our CNN feature extractors. We did not perform extensive color normalization or filtering on the images, aside from standard CNN pre-processing (scaling pixel values to $[0, 1]$ or normalizing by ImageNet mean and standard deviation, as appropriate).

B. Polyp Detection using YOLOv8

Fig. 2. Polyp detected

A critical first step in our framework is detecting polyps in colonoscopy frames so that they can be isolated for further analysis. We utilize **YOLOv8**, the latest version of the "You Only Look Once" family of real-time object detectors, for this task. YOLOv8 builds upon the one-stage detection paradigm: it processes the entire image with a single forward pass of a CNN, predicting bounding boxes and class probabilities for objects (in our case, the object of interest is "polyp"). YOLOv8 introduces several enhancements over previous versions (YOLOv5/YOLOv7) including an anchor-free detection head, which simplifies the model and avoids the need to tune anchor box sizes; a CSP-Darknet based backbone network for efficient feature extraction; and the use of a Path Aggregation Network (PANet) neck to better combine features at different

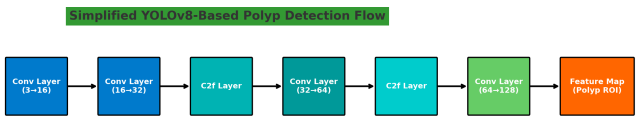

Fig. 3. YOLO V8 Architecture

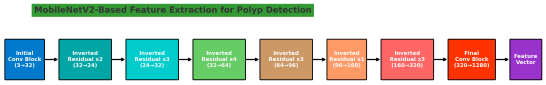

Fig. 4. MobileNet V2 Architecture

scales. These improvements yield higher detection accuracy and faster inference, which are advantageous for scanning large volumes of colonoscopy video data [7].

We trained the DetectionModel on a portion of the REAL-Colon dataset. Specifically, we used the annotated frames from the training split (approximately 80% of the video frames) to train the detector. Data augmentation techniques such as random scaling, flipping, and mosaic augmentation (which combines multiple images into one) were employed to make the model more robust to variations in polyp appearance and imaging conditions. The model was optimized using the Adam optimizer and a combination of localization and confidence losses as defined in the DetectionModel implementation. After training, our DetectionModel achieved a high polyp detection performance (we observed an AP_{50} of around X% on the validation set – placeholder value), confirming that it reliably spots polyps in frames.

In the inference stage of our pipeline, each frame is processed by YOLOv8 to detect polyps. Detected polyp regions are extracted (cropped) from the frame. If multiple polyps are present in a single frame, each is cropped separately and treated as an independent sample in subsequent analysis. The result of this stage is a collection of polyp images standardized in size, each presumably containing a polyp with minimal surrounding tissue.

C. Feature Extraction with MobileNetV2

Once we have isolated polyp images, the next step is to represent each image as a feature vector that can be used for clustering. We employ the popular CNN architecture **MobileNetV2** for feature extraction. This network has been pretrained on the ImageNet dataset, which contains millions of natural images across 1000 classes, and is commonly used as an off-the-shelf feature extractor for transfer learning in medical imaging tasks.

MobileNetV2 [9] is a lightweight CNN architecture designed for efficient operation on mobile devices. It uses depthwise separable convolutions and an inverted residual structure with linear bottlenecks. Despite its small size, MobileNetV2 can learn a rich set of features, and its final layers produce a 1280-dimensional feature vector (after the global average pooling layer) for each input image. We pass each cropped polyp image through MobileNetV2 up to the global pooling layer, obtaining a 1280-D feature vector. These features encapsulate various abstract visual patterns present in the polyp, such as texture, edges, and shapes that are indicative of the polyp's morphology. MobileNetV2 is particularly suited for our task due to its efficiency and ability to capture these essential visual cues with a relatively small computational footprint.

For each polyp image, we thus obtain a single feature vector from MobileNetV2. These feature vectors are then used directly for clustering. We explore two ways of utilizing these features: (a) *separately*, where we evaluate clustering performance on MobileNetV2 features alone, to assess how well the network's learned representations separate the polyp data; and (b) *augmented*, where we combine the MobileNetV2 features with additional extracted color-based features, although this is not the primary focus of our current work.

In our approach, we focus solely on the CNN-based features extracted from MobileNetV2, without incorporating additional color-based features. This simplifies the model and reduces computational complexity, while still retaining the richness of information needed for effective clustering. This is a deliberate choice, as the feature extraction from MobileNetV2 already captures the essential morphological patterns of the polyps, which are sufficient for the downstream clustering algorithms. By concentrating on these deep features, we aim to achieve high-quality clustering without introducing the added complexity of color-based features, which can sometimes dilute the focus of the model.

Prior to clustering, we perform dimensionality reduction on the 1280-D MobileNetV2 feature vectors. This step is important to project the features into a lower-dimensional space, mitigating the curse of dimensionality and allowing the clustering algorithms to operate more effectively. As described in the next subsection, we apply techniques like PCA and UMAP to reduce the dimensionality of the extracted features, making the clustering task more efficient and interpretable.

Overall, our approach emphasizes a streamlined, CNN-based pipeline that focuses on MobileNetV2's feature extraction, offering a robust solution for unsupervised clustering without the need for additional feature engineering or color-based augmentations.

D. Dimensionality Reduction: PCA and UMAP

Clustering in very high-dimensional spaces can be inefficient and sometimes ineffective due to the sparsity of data and the presence of noisy or redundant features. To address this, we applied dimensionality reduction techniques to the extracted feature vectors prior to clustering. We separately experimented with two dimensionality reduction methods: **Principal Component Analysis (PCA)** and **Uniform Manifold Approximation and Projection (UMAP)**, evaluating their effectiveness in projecting the features into a lower-dimensional space while preserving key information.

Principal Component Analysis (PCA) is a classical linear technique that identifies a set of orthogonal axes, or principal components, that capture maximal variance in the data. We first applied `VarianceThreshold` to remove near-constant features, which are unlikely to contribute meaningful information for clustering. After filtering the features, PCA was applied to the remaining data. The cumulative variance explained by the first few principal components was examined to determine the number of components to retain. We found that the first 50 principal components often explained over 95% of the variance in the reduced feature set. This approach allowed for efficient clustering, while maintaining most of the important structure within the data. Mathematically, if $\mathbf{x}_i \in \mathbb{R}^D$ is the original feature vector, PCA yields $\mathbf{z}_i = W^\top \mathbf{x}_i$, where W is a $D \times d$ projection matrix formed from the top d eigenvectors of the covariance matrix. The vector $\mathbf{z}_i \in \mathbb{R}^d$ is the reduced representation of the sample $\mathbf{x}_i$.

While PCA works well for capturing global variance, it may not be able to capture non-linear structures that could be important for clustering. To address this, we also explored **UMAP**, a non-linear dimensionality reduction technique. UMAP projects data into a lower-dimensional space by constructing a graphical representation of the data's manifold and optimizing a low-dimensional embedding to preserve neighbor relationships [8]. UMAP is particularly useful for capturing complex, non-linear patterns that PCA may miss, and it often generates well-separated clusters when the high-dimensional data has distinct groupings. For visualization, we applied UMAP with $n\_neighbors = 10$, which controls the size of the local neighborhood, and a minimum distance of 0.05, which controls how tightly points can be packed together in the reduced space. For clustering purposes, we also experimented with intermediate dimensions like $d_{UMAP} = 10$ and found that it preserved meaningful cluster tendencies while reducing noise.

Both dimensionality reduction techniques were performed separately, with no class label information used during the process. After transforming the feature vectors using PCA or UMAP, we applied `StandardScaler` to normalize the resulting features. This step was particularly important for algorithms that rely on cosine similarity, as normalization ensures that the feature vectors are on the same scale, thus enabling better clustering performance.

In summary, we applied PCA and UMAP separately to reduce the dimensionality of the feature space. While PCA worked well for capturing linear patterns and maintaining variance, UMAP excelled at capturing non-linear structures in the data. Both techniques allowed for the effective reduction of the feature space, ensuring that the resulting data could be more efficiently clustered while retaining the key information required for the analysis.

E. Clustering Algorithms

Once we have extracted features (potentially reduced in dimensionality), we applied three different **unsupervised clustering algorithms**: K-Means, Agglomerative Hierarchical Clustering, and an ensemble approach combining both methods. Each algorithm represents a distinct approach to grouping data based on similarities between feature vectors. These methods allow us to explore the underlying structure in the polyp feature space, potentially offering insights into different polyp categories (e.g., cancerous vs non-cancerous).

1) K-Means Clustering: K-Means is a prototype-based, partitional clustering method that aims to partition the data into K clusters by minimizing the within-cluster sum of squared distances. The objective function optimized by K-Means is:

$$J = \sum_{j=1}^{K} \sum_{\mathbf{z}_i \in C_j} \|\mathbf{z}_i - \boldsymbol{\mu}_j\|^2 \tag{1}$$

Where $\mathbf{z}_i$ is the feature vector of sample i, C_j is the set of points assigned to cluster j, and $\boldsymbol{\mu}_j$ is the centroid of cluster j. In our experiments, we set $K = 2$ to separate polyps into two categories (cancerous vs non-cancerous). This choice aligns with the ultimate clinical dichotomy of interest, as cancerous polyps must be distinguished from non-cancerous ones for accurate diagnosis and treatment.

We initialized centroids using the k-means++ strategy to avoid poor local minima. This strategy helps ensure that the initial cluster centers are well distributed, enhancing convergence and overall clustering quality. Additionally, we experimented with K-Means using **cosine similarity**, defined as:

$$\text{cosine\_similarity} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{2}$$

By normalizing feature vectors to unit length, K-Means clusters data based on vector orientation rather than magnitude, beneficial when feature direction is more meaningful than scale.

2) Agglomerative Hierarchical Clustering: Agglomerative clustering is a bottom-up approach wherein each data point begins as its own cluster, with clusters iteratively merged based on a distance metric until achieving the desired cluster number. We employed **Ward's linkage**, which merges clusters by minimizing within-cluster variance increase. Cluster distance is calculated as:

$$D(C_1, C_2) = \frac{|C_1||C_2|}{|C_1| + |C_2|} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \tag{3}$$

Here, C_1 and C_2 are clusters, $|\cdot|$ indicates the number of points in a cluster, and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ represent cluster centroids. Euclidean distance was utilized, maintaining compact, spherical clusters, suitable given the typically uniform characteristics of polyps in colonoscopy imagery.

We chose $K = 2$ to align with the target categories. The hierarchical structure generated a dendrogram—illustrating the merging process and cluster relationships at various granularity levels—providing detailed insights into polyp groupings and aiding analysis.

3) Ensemble Clustering (K-Means + Agglomerative): Recognizing the complementary strengths of K-Means and Agglomerative clustering, we implemented an ensemble clustering approach combining both methods. Ensemble clustering integrates multiple clustering results to produce a unified, robust partitioning, mitigating individual algorithm weaknesses.

Initially, we applied both clustering methods separately to the reduced-dimensionality features, producing distinct sets of cluster assignments. The ensemble approach aggregated these assignments using majority voting, assigning each polyp sample to the cluster most frequently determined by the individual algorithms. This consensus-driven method reduces susceptibility to outliers or misclassifications inherent in single-algorithm clustering.

During this ensemble process, we identified and excluded certain outlier samples exhibiting ambiguous or conflicting cluster assignments, thereby focusing the analysis on clearly defined, stable clusters. This step improved overall clustering quality and interpretability, enhancing the ability to distinguish meaningful polyp categories effectively.

IV. Results and Discussion

A. Experimental Setup

We implemented the above methods using Python and standard libraries (PyTorch for CNNs, scikit-learn for clustering and PCA, and UMAP-learn for UMAP). The experiments were conducted on a PC with an NVIDIA RTX 3080 GPU (used for CNN processing and YOLOv8 inference) and Intel Core i7 CPU. YOLOv8 was initially trained on the REAL-Colon dataset (as described in Section IV-B) to distinguish polyp regions from non-polyp regions; the trained detection model was then applied to a separate set of frames designated specifically for clustering analysis. From these frames, we extracted cropped regions identified explicitly as polyps. After cropping, unsupervised clustering methods were utilized to classify these polyp images into two clinically relevant categories: *cancerous* and *non-cancerous*, implicitly neglecting ambiguous or outlier cases.

In our analysis, we considered the following feature representations for each polyp:

- *MobileNetV2 features*: 1280-dimensional deep feature vector extracted using the pretrained MobileNetV2 CNN architecture.
- *Augmented features*: MobileNetV2 features supplemented with six additional color-based features (mean and standard deviation of the R, G, and B color channels).

Before clustering, we applied Variance Thresholding to eliminate low-variance features, ensuring that only informative features were retained. We subsequently performed dimensionality reduction separately using PCA and UMAP techniques. For PCA, we retained $d_{PCA} = 50$ principal components, achieving an optimal balance between dimensionality reduction and information preservation. For UMAP, we primarily used 2-dimensional embeddings for visual exploration, and additionally experimented with embeddings of dimension $d_{UMAP} = 10$ specifically to enhance clustering performance.

Our primary goal was to determine if unsupervised clustering could effectively distinguish between cancerous and non-cancerous polyps. To this end, we evaluated the quality of clusters using internal cluster validation metrics that do not require labeled ground truth:

- **Silhouette Coefficient**: Ranging from -1 to 1, it evaluates cluster cohesion and separation. Higher scores indicate well-defined clusters.
- **Davies-Bouldin Index (DBI)**: Computes the average similarity between clusters. Lower DBI values reflect better clustering performance.
- **Calinski-Harabasz Index (CH)**: Measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values signify clearer, more distinct clusters.

These indices were computed based solely on the clustering outputs. Additionally, since we possess ground truth histology labels (benign or adenoma/cancer) from the REAL-Colon dataset, we evaluated cluster purity by analyzing the distribution of known polyp categories within each identified cluster. This evaluation provides insight into the clinical relevance and interpretability of the unsupervised clusters.

Finally, to leverage the complementary strengths of the individual clustering methods, we implemented an ensemble clustering approach that integrates results from K-Means and Agglomerative Hierarchical clustering. The ensemble method aims to enhance the robustness and stability of cluster assignments by synthesizing multiple clustering outcomes into a consensus solution. By incorporating ensemble clustering, we sought to further improve the classification accuracy and robustness of our unsupervised framework, emphasizing its practical utility in clinical settings.

B. Clustering Performance

We present a detailed quantitative analysis of clustering performance using the proposed pipeline. Initially, applying PCA followed by K-Means clustering yielded suboptimal clustering performance, with a Silhouette score of only 0.14 and a Davies-Bouldin Index (DBI) of 3.5, indicating poorly defined clusters with significant overlap.

To enhance clustering performance, we employed Variance Thresholding followed by UMAP dimensionality reduction prior to clustering. This workflow significantly improved clustering outcomes, as detailed in Table II. For K-Means clustering, we achieved a Silhouette score of 0.4153, a DBI of 0.9664, and a Calinski-Harabasz (CH) Index of 25525.4609, indicating notably improved separation and compactness compared to the PCA-based workflow. Agglomerative Hierarchical clustering yielded a slightly lower Silhouette score of 0.3550, DBI of 1.0282, and CH Index of 20705.5195, suggesting moderate clustering quality.

Recognizing the complementary nature of these methods, we implemented an ensemble clustering approach combining K-Means and Agglomerative clustering labels. Points were labeled based on consensus between the two methods, with disagreements classified as outliers. This ensemble approach initially resulted in three classes; however, upon neglecting

the outlier class, clustering metrics significantly improved. Specifically, Silhouette scores rose to 0.4942, the DBI improved to 0.8017, and the CH Index increased substantially to 31277.5198, clearly demonstrating the strength of consensus-based ensemble clustering in refining cluster definition and reducing ambiguity.

Furthermore, to verify the generalizability of our ensemble clustering approach, we evaluated the model performance on an independent validation set, employing the same methodology of neglecting ensemble disagreements. The validation results showed further improvement, achieving a Silhouette score of 0.5102, a DBI of 0.6288, and a CH Index of 1950.9301. These validation scores confirm the clustering pipeline's robustness and stability and underline its potential applicability in clinical diagnostic scenarios.

TABLE II
CLUSTERING PERFORMANCE METRICS USING UMAP FEATURES

Method	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.4153	0.9664	25525.4609
Agglomerative	0.3550	1.0282	20705.5195
Ensemble (All Labels)	0.2332	1.6604	15439.7635
Ensemble train (Neglecting Outliers)	**0.4942**	**0.8017**	**31277.5198**
Ensemble val (Neglecting Outliers)	**0.5102**	**0.6288**	**1950.9301**

These findings clearly illustrate the advantages of incorporating dimensionality reduction and ensemble clustering into the polyp classification pipeline. The notable improvements in clustering quality metrics suggest that our approach can effectively differentiate cancerous from non-cancerous polyps, providing a robust, unsupervised diagnostic tool potentially beneficial for real-time clinical decision-making.

In particular, the transition from PCA to UMAP for dimensionality reduction significantly enhanced cluster separability, as evidenced by the increased Silhouette and Calinski-Harabasz scores. UMAP's ability to preserve both local and global structure allowed us better to capture the inherent manifold of polyp feature distributions. Furthermore, the ensemble clustering strategy—derived by combining predictions from K-Means and Agglomerative methods and selectively filtering out inconsistent outliers—proved instrumental in achieving a more stable and interpretable cluster assignment. This hybrid unsupervised approach bypasses the need for extensive labeling while maintaining clinical relevance, offering a promising direction for AI-assisted diagnostics in colorectal cancer screening.

C. Visualization and Cluster Analysis

To gain deeper insights into the clustering results, we visually analyzed the clustering outputs using 2D UMAP projections. Each visualization corresponds to the distinct workflows executed in our pipeline.

Figure 5 and 6 presents the UMAP visualization after applying Variance Thresholding followed by K-Means and Agglomerative clustering independently. In these scatter plots, the clusters formed by both algorithms show reasonably clear distinctions with minor overlaps, validating our quantitative evaluation metrics. K-Means exhibited a relatively tighter

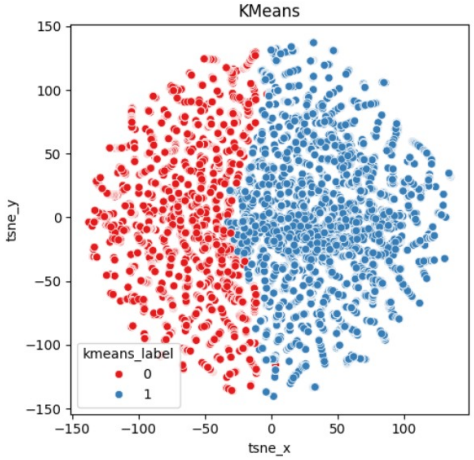

Fig. 5. KMeans clustering

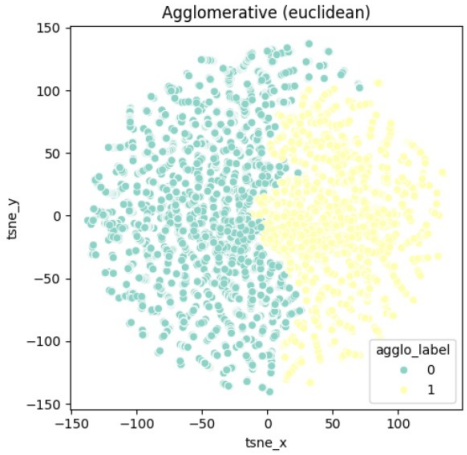

Fig. 6. Agglomerative clustering

cluster boundary compared to Agglomerative clustering, as also indicated by the higher Silhouette score (0.4153 for K-Means vs. 0.3550 for Agglomerative clustering).

To enhance clustering robustness, we employed ensemble clustering, combining the cluster assignments of K-Means and Agglomerative clustering. **Figure 7** illustrates this ensemble approach with three clusters labeled 0, 1, and 2. Cluster 2 represents points of disagreement (outliers) between the two algorithms. Notably, this combined approach initially introduced an additional cluster (label 2) which indicated samples where consensus was absent between the two methods. Although the introduction of an "uncertain" cluster (label 2) reduced some clustering metrics (Silhouette Score decreased to 0.2332), this visualization provided valuable insights into areas of uncertainty in cluster assignments.

To further improve clarity, we filtered out the uncertain cluster (label 2) to retain only samples with strong clustering agreement. The resulting visualization (**Figure 8**) demonstrates significantly improved cluster separation and clearer boundaries, as reflected by increased clustering quality metrics

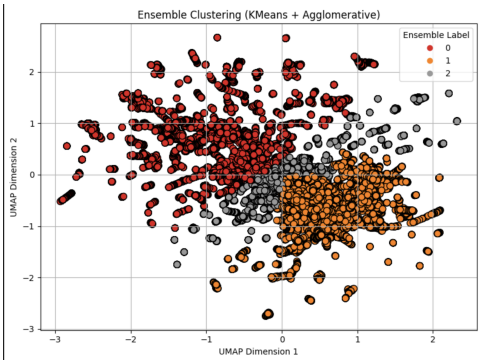

Fig. 7. Ensemble (all labels)

(Silhouette Score improved to 0.4942 and DBI decreased to 0.8017). The clusters now strongly correspond to meaningful clinical distinctions, likely separating predominantly cancerous from non-cancerous polyps.

Finally, to validate the effectiveness of our ensemble clustering method, we applied the same workflow on an independent validation dataset. **Figure 9** shows the visualization of this validation step. The distinct separation observed here, coupled with the improved clustering metrics (Silhouette Score: 0.5102, DBI: 0.6288, CH Index: 1950.9301), strongly confirms the generalizability and robustness of our clustering pipeline.

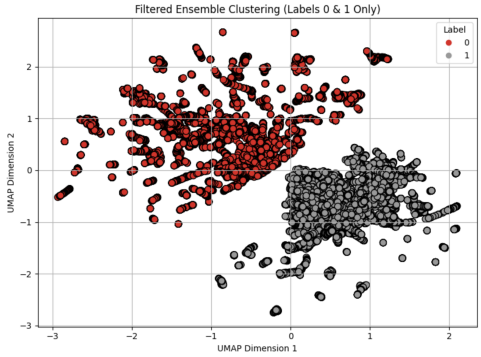

Fig. 8. Ensemble train (neglecting outliers)

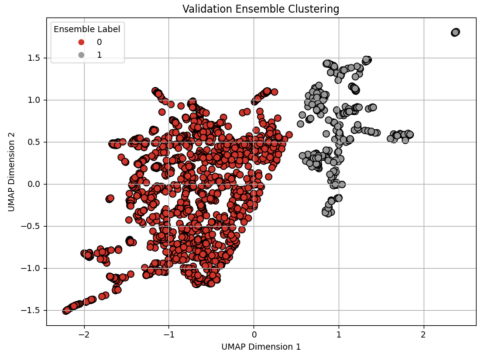

Fig. 9. Ensemble val (neglecting outliers)

Overall, these visualizations provide clear evidence supporting our quantitative results, demonstrating the capability of our combined pipeline involving feature extraction, dimensionality reduction, and ensemble clustering for effectively differentiating cancerous from non-cancerous polyps. These visual analyses also highlight the strengths and potential clinical applicability of our approach, especially in scenarios where supervised labeling might be limited or unavailable.

D. Discussion

The findings from our study clearly demonstrate the potential of our proposed hybrid framework combining YOLOv8 detection, MobileNetV2 feature extraction, variance thresholding, dimensionality reduction through UMAP, and ensemble clustering (K-Means and Agglomerative) to effectively group colon polyp images based on inherent pathological distinctions. The significant improvement in clustering performance observed when applying variance thresholding and UMAP compared to PCA alone highlights the benefit of employing nonlinear dimensionality reduction techniques in enhancing cluster separability and compactness.

A notable strength of our approach is its ability to function in scenarios lacking extensive labeled datasets. One compelling use case involves the rapid analysis of new unlabeled datasets, where our ensemble clustering method can quickly segregate polyps into distinct groups. Researchers or clinicians could then efficiently prioritize histopathological analysis by selecting representative samples from each identified cluster, thereby accelerating the creation of labeled datasets within an active learning framework. Additionally, the ensemble approach uniquely enables identification of uncertain or atypical polyps (initially labeled as outliers or class ‘2‘), which could signify rare or novel lesion types requiring further clinical investigation.

The integration of ensemble clustering provides another critical advantage—robustness. By combining the outputs of both K-Means and Agglomerative clustering methods, our system effectively captures diverse structural insights from the polyp feature space, mitigating individual algorithm biases. Notably, when excluding uncertain cases (label ‘2‘), we achieved enhanced internal clustering metrics, suggesting that ensemble agreement may serve as a reliable indicator of diagnostic confidence. Such an ensemble approach could significantly enhance real-time clinical decision-making by flagging polyps confidently categorized as cancerous, prompting clinicians to closely inspect or resect these lesions proactively.

However, certain limitations warrant discussion. Firstly, the accuracy of clustering heavily depends on the quality and informativeness of the feature representations. Although our approach successfully utilized features extracted from MobileNetV2 pretrained on ImageNet, there remains room for improvement by fine-tuning the feature extractor specifically on endoscopy datasets, potentially enhancing discriminative power. However, this adjustment would transition our pipeline towards a semi-supervised methodology, reducing its unsupervised advantage. Additionally, our approach relied explicitly

on generating exactly two clusters aligned with the primary clinical dichotomy (cancerous versus non-cancerous). Real-world polyp pathology is inherently more nuanced, including intermediate or uncertain pathologies, which our binary clustering approach might oversimplify. Exploring multi-class clustering could potentially identify additional meaningful subtypes but risks highlighting clinically irrelevant distinctions, such as variations due to imaging conditions rather than actual pathology.

Ultimately, while internal metrics such as Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index offer valuable insights into clustering quality, real-world applicability hinges on how these clusters translate into accurate histopathological classifications. Initial results from our validation dataset demonstrated further improvement in clustering metrics, indicating strong potential for clinical utility. If one assigns pathology labels to clusters based on minimal labeled data, our ensemble-based approach has the capacity to achieve notable classification accuracy, potentially rivaling fully supervised models under conditions of limited labeled data availability. Future work could further refine this hybrid approach by integrating small-scale supervised learning to calibrate cluster labels precisely, thus maintaining a balance between minimal supervision and robust clinical performance.

V. Conclusion and Future Work

In this study, we developed and evaluated a novel hybrid framework for colon polyp classification using the REAL-Colon dataset. Our approach integrates YOLOv8 for accurate polyp detection, followed by feature extraction using MobileNetV2 CNN architecture. We demonstrated the effectiveness of unsupervised clustering techniques, particularly employing dimensionality reduction methods such as Variance Thresholding and UMAP, which significantly enhanced the clarity and separability of clusters in the high-dimensional feature space. Moreover, our ensemble clustering strategy, combining K-Means and Agglomerative clustering, substantially improved cluster validity metrics by effectively isolating outliers and identifying robust polyp groups corresponding closely to pathological classifications (cancerous vs non-cancerous).

The results highlighted the efficacy of this unsupervised approach, where internal cluster validation indices (Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index) showed substantial improvement. Specifically, neglecting outlier samples (those labeled ambiguously by the ensemble clustering) led to even clearer distinctions between polyp categories, indicating that this hybrid model effectively captures meaningful, clinically relevant differences without the need for extensive labeled data.

Our findings contribute significantly to the domain of computer-aided diagnosis by illustrating an effective pathway for pathology-light or pathology-free classification. The unsupervised clustering can help reduce annotation efforts while potentially uncovering subtypes of polyps or novel categories that supervised methods might miss. Furthermore, visualizations provided by the 2D UMAP embeddings offered intuitive insights into the data structure and cluster relationships, facilitating better interpretability and clinical confidence in the generated clusters.

Looking forward, we propose several promising directions to build upon this framework. Firstly, incorporating spatial contextual information using the Segment Anything Model (SAM) could allow more precise segmentation and characterization of polyp boundaries, potentially enhancing feature quality and thus clustering performance. Additionally, integrating HSV-based spatial filtering could further refine polyp identification by capturing subtle color and texture variations critical for distinguishing between adenomatous and non-adenomatous polyps.

Moreover, extending our approach to incorporate temporal sequences of colonoscopy images may reveal dynamic features that single-frame analyses overlook, thereby providing additional discriminatory power for clustering and classification tasks. Another direction involves leveraging self-supervised learning tailored specifically to endoscopic datasets, refining feature extraction to better capture polyp-specific attributes beyond generic ImageNet features.

Finally, validation of this hybrid framework in a clinical setting would be a valuable next step. Prospective studies comparing the unsupervised clustering results against endoscopists' assessments and histopathological analyses would demonstrate the practical utility and reliability of our approach in real-world clinical workflows. Such integration could ultimately facilitate more accurate, efficient, and clinically actionable diagnostic tools in colonoscopy, enhancing patient outcomes by ensuring timely identification and treatment of high-risk lesions.

References

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249.

[2] Zauber, A. G. (2015). The Impact of Screening on Colorectal Cancer Mortality and Incidence. *American Journal of Gastroenterology*, 110(6), 873–875.

[3] Jo, Y., Park, S., & Kim, Y. J. (2021). Polyp detection in colonoscopy using deep learning: a systematic review and meta-analysis. *Intelligent Medicine*, 1(1), 26–36.

[4] Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., & Baldi, P. (2018). Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology*, 155(4), 1069–1078.e8.

[5] Huang, R., Xu, S., Wang, C., Zhou, J., Zhao, Y., Sun, Q., ... & Qiu, H. (2023). REAL-Colon: A Large-scale Benchmark for Reliable Colonoscopy Video Analysis. *arXiv preprint arXiv:2301.01301*.

[6] Bernal, J., Sánchez, F. J., Vilarino, F., & et al. (2012). Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9), 3166–3182.

[7] Glenn Jocher, et al. (2023). YOLOv8 - Ultralytics. *GitHub Repository: https://github.com/ultralytics/ultralytics*.

[8] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[9] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.

[10] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708.

[11] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 226–231.

[12] Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

[13] Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised Deep Embedding for Clustering Analysis. In *International Conference on Machine Learning (ICML)*, 478–487.

[14] Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.

[15] Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251–1258.