International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

AGFC - Augmenting Image Analysis: Gradient Magnitude from a Smoothed Image for Improved Feature Detection in Colorectal Imagery

Madduri Deepika*[1], Dr. D. Thiyagarajan[2], Dr. D. Murali[3]

Submitted:14/03/2024 **Revised:** 29/04/2024 **Accepted:** 06/05/2024

Abstract: This research presents a comprehensive framework for the processing and classification of multi-modal colorectal images, leveraging an extensive array of data augmentation, neural network models, and advanced techniques. The multi-level classification pipeline commences with a Sequential Convolutional Neural Network (SCNN) and progresses to the subsequent stage, featuring an abnormal tissue detection module incorporating excess object removal and transformers. The architecture further integrates a hybrid Convolutional Neural Network (HCNN), encompassing a Vision Transformer (ViT), a custom cross-modality transformer, a traditional CNN, a Multilayer Perception (MLP), and a combined model. The apex of this approach materializes in a final multi-modal classifier, validating testing images and executing classification tasks. This framework not only showcases a sophisticated and effective strategy for multi-modal colorectal image processing but also exhibits the potential to augment the precision and generalization of Colorectal Cancer (CRC) risk assessments. The incorporation of diverse imaging modalities and advanced neural network architectures positions this method as a robust tool for refining the accuracy of CRC risk predictions in clinical applications.

Keywords: *Colorectal Cancer, HCNN, Multilayer Perception, MLP, SCNN.*

1. Introduction

This method explores the intersection of computer science and medicine, presenting a dynamic fusion that holds immense potential for revolutionizing healthcare practices. The initial discourse focuses on the role of Machine Learning (ML), a standout in computer science, in addressing medical challenges. It unfolds the versatility of ML applications in pathology, from disease identification to the creation of intelligent systems capable of recommending traditional medications based on patient symptoms [8].Shifting to the realm of cancer risk, the research explores a strong birth cohort effect, signaling population-level changes in behavioral factors influencing cancer susceptibility. Despite coverage by major insurers, the slow implementation of screening in community settings is highlighted. Moreover, it accentuates the untapped potential of identifying high-risk families, even in the absence of a genetic syndrome, as a strategic approach to mitigating the cancer burden through early screening.

The narrative then transitions to recent discussions surrounding CRC diagnosis, emphasizing challenges faced in conventional diagnostic methods and innovative solutions on the horizon. The traditional gold standard of CRC diagnosis via electronic colonoscopy is examined, with an acknowledgment of the challenges of endoscopic disinfection standards [3]. Lifestyle factors, dietary patterns, and the westernization of lifestyle are recognized as influencers of CRC morbidity, pointing toward the need for novel diagnostic approaches [15].Expanding the scope, the proposal incorporates discussions on microbiome-based approaches for CRC screening, underscoring the impact of biomarker selection identification algorithms on massive gut microbiome data [9]. Furthermore, it touches upon the challenges faced in developing Deep Learning (DL) algorithms for medical image analysis, emphasizing the need for well-organized and labeled training data and defined rules for algorithm comparison [13].

Our approach integrates an SCNN with an abnormal tissue detection module, leveraging transformers. The architecture extends to an HCNN, encompassing a ViT, a cross-modality transformer, a traditional CNN, MLP, and a combined model. The final multi-modal classifier enhances precision and generalization in CRC risk assessments, showcasing the efficacy of advanced neural network architectures and diverse imaging modalities as shown in Figure1.

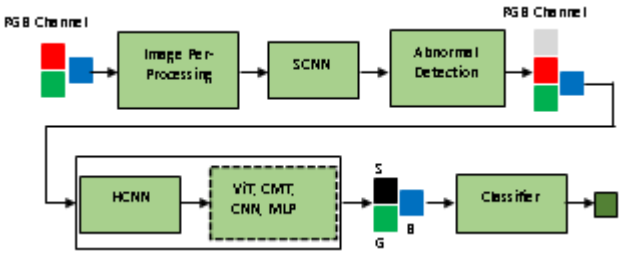


Figure 1.Diverse imaging modalities employed in the proposed

The structure of the document unfolds in the following manner: Section 2 probes into related works, providing an overview of previous studies related to the proposed method. Section 3 comprehensively explores the proposed

method and its development flows. Moving forward, Section 4 examines the results and discussions on both existing and proposed methods. The concluding section summarizes the proposed method, highlights its implications, and outlines avenues for future research in Section 5.

2. Related works

The realm of medical imaging has undergone a transformative evolution, propelled by the advancements in ML methodologies. DL strategies have enabled machines to decipher high-dimensional data, encompassing diverse formats such as images, multimodal pathology scans, and video files. Particularly adept in handling biological images, numerous supervised machine-learning techniques have been developed [5]. In Low- and Middle-Income Countries (LMICs), where achieving optimal cancer care requires cost-effective interventions, leveraging ML in healthcare planning becomes pivotal. This ensures an equitable distribution of healthcare resources, addressing health disparities on a global scale. Employing a health continuum approach becomes imperative, guiding public awareness campaigns about the benefits of screening programs and the significance of recognizing early signs of cancer [1].

In the quest for advancing early cancer diagnosis, we underscore the importance of innovative liquid biopsy-based tests. Our comprehensive review delves into the current landscape of liquid biopsy modalities, exploring their role in early cancer diagnosis and monitoring. Emphasizing both technical and clinical challenges intrinsic to the development of clinically relevant liquid biopsy assays, we underscore the necessity for adopting best practices. Establishing these practices becomes instrumental in navigating the biomarker discovery pipeline, ultimately enhancing the translational potential of liquid biopsy findings [7]. This holistic approach seeks to propel the integration of cutting-edge technologies into clinical practice, fostering a new era of precision medicine in cancer care.

Addressing the intricate differences in healthcare organization, delivery, resources, infrastructure, and social norms between LMICs and High-Income Countries (HICs) demands a systematic, system-strengthening approach. This approach is essential not only to bridge the gaps but also to engage eligible screening populations effectively. The design and delivery of screening interventions must intricately consider the complex implementation considerations inherent in diverse global healthcare landscapes [2].

The integration of genomic data has empowered physicians and healthcare decision-makers to delve deeper into understanding patients and their responses to therapy. This integration has spurred the application of ML and DL to tackle complex challenges in cancer research. Tasks include the creation of cancer risk-prediction models aimed at identifying individuals at a heightened risk of developing cancer and studying the disease's progression [16].

CRC, largely preventable through the avoidance of modifiable risk factors and early detection, has witnessed a surge in early-onset cases (EOCRC). While it was traditionally linked to hereditary syndromes, the contemporary rise is attributed to widespread inactive lifestyles and unhealthy eating habits globally [6]. The expansion of screening colonoscopy programs holds the potential to increase early CRC diagnoses. Consequently, the development of additional reporting categories that offer nuanced prognosis stratification for these patients becomes imperative. This review delves into novel concepts and challenges associated with the pathological assessment and reporting of CRC [4].

This comprehensive exploration aims to pave the way for innovative strategies in the diagnosis, prevention, and management of CRC on a global scale. In recent times, ML has emerged as a powerful tool for cancer prognosis and prediction, aligning seamlessly with the evolving paradigm of personalized and predictive medicine. This transformative strategy has influenced the landscape of cancer development and therapy, steering them toward more tailored and effective approaches. A crucial aspect of this application is the ability of ML algorithms to discern significant patterns within vast datasets. This capability is harnessed to develop prediction models that anticipate the onset and potential cure of cancer [10].

The proposed methodology Automated Gastrointestinal Feature Classification (AGFC) involves the creation of a robust machine-learning classification model. It integrates feature extraction methods, explores multi-view medical image registration, employs fusion algorithms, and evaluates performance using well-defined indicators. The overarching goal of this research is to make a meaningful contribution to the field by presenting a reliable and accurate solution for the detection of lung cancer. Such advancements are poised to significantly enhance patient outcomes, marking a substantial stride in the realm of medical applications within this domain [11].

This innovative approach holds the promise of not only improving the accuracy of cancer detection but also revolutionizing the landscape of medical practices in cancer care. The conventional gold standard for diagnosing CRC involves obtaining materials through electronic colonoscopy for pathological confirmation. Additionally, endoscopy emerges as an effective method for treating patients experiencing gastrointestinal hemorrhage or ileus without peritonitis. However, challenges arise as the

previous endoscopic disinfection standard may not ensure the inactivation of new coronavirus, posing risks of doctor-patient and patient-patient cross infections [3].

The surge in CRC morbidity is intricately linked to lifestyle choices, body fatness, and dietary patterns. Convincing evidence suggests that physical activity offers protective benefits, while increased consumption of red and processed meat, as well as alcoholic drinks, escalates the risk of developing the disease. Societal progress and economic development, while improving socioeconomic conditions, also trigger a shift in dietary patterns, often referred to as the Westernization of lifestyle [15].

The mini-review explores the impact of the biomarker selection identification algorithm on extensive gut microbiome data in altering the performance of CRC diagnosis. It provides a comprehensive summary of current microbiome-based approaches for CRC screening, encompassing experimental design, markers selection, and identification methods. The review also proposes potential solutions to enhance CRC detection and prediction [9].

Despite the promise of DL algorithms in medical image analysis, significant challenges persist, including the scarcity of well-organized and labeled training data, a lack of benchmark and test data, and a need for clearly defined rules for comparing algorithms. This becomes crucial as systems with high false-positive rates may diminish radiologist sensitivity. Moreover, limited access to previously developed algorithms for comparison adds to the complexity [13].

The effectiveness and efficiency of ML solutions hinge on the nature and characteristics of data and the performance of learning algorithms. Within the realm of ML algorithms, various techniques such as classification analysis, regression, data clustering, feature engineering, dimensionality reduction, association rule learning, and reinforcement learning play a vital role in constructing data-driven systems [12].

This comprehensive exploration illuminates the multifaceted landscape of CRC diagnosis, addressing challenges and offering potential breakthroughs in the ongoing quest for improved cancer detection and prediction. A potential resolution to the challenges faced in medicine and healthcare has emerged from an unexpected source – the realm of computer science. Remarkably progressive in comparison to other scientific and technological fields, computer science has provided a unique perspective. The strides made in ML, a subset of computer science, offer a wide range of applications in pathology, spanning from disease identification to the development of intelligent systems capable of recommending traditional medications based on a patient's symptoms [8].

Within the context of cancer risk, a robust birth cohort effect points toward population-level changes in behavioral factors. Despite being covered by major public and private insurers, the implementation of screening has been sluggish in community settings. However, identifying high-risk families, even in the absence of a genetic syndrome, presents a significant opportunity to alleviate the burden of cancer through early screening initiatives [14]. This innovative convergence of medicine and computer science holds promise for advancing healthcare practices, introducing intelligent solutions, and redefining approaches to cancer prevention and identification.

3. Proposed Methodology

3.1. Categorization of Images

Colorectal input images have been squarely collected and organized into seven distinct classes, each housed in its dedicated folder. These classes encompass Tumor, Stroma, Complex, Lympho, Debris, Mucosa, and Adipose, providing a comprehensive representation of the diverse elements present in colorectal samples. This thorough categorization enhances the efficiency of image analysis and facilitates a nuanced understanding of the various components within colorectal tissues, contributing to advancements in research, diagnostics, and medical interventions related to colorectal health.

The ImageDataGenerator is employed for performing data augmentation on training images. This tool in ML and computer vision enhances the robustness of the training dataset by applying various transformations to the images, such as rotation, scaling, and horizontal flipping. Data augmentation is instrumental in preventing overfitting and improving the model's generalization capabilities, enabling it to better handle diverse and real-world scenarios. By introducing variations in the training data through the ImageDataGenerator, the model becomes more adept at recognizing patterns and features, ultimately leading to more effective and accurate predictions during the training process as shown in Figure2.

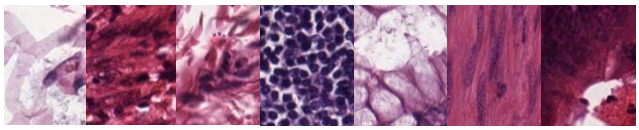


Figure 2. Training dataset

The ImageDataGenerator plays a significant role in augmenting training images by incorporating multiple transformations. This includes rescaling pixel values to a standardized range, applying shear transformations as shown in Figure 3to introduce deformation, implementing zooming to simulate variations in perspective, and incorporating random horizontal flipping for added

diversity.

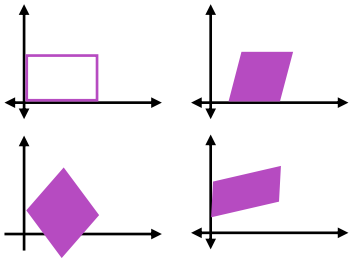


Figure 3. Shear transformation

These augmentation techniques contribute to a more robust and versatile training dataset, enabling ML models to better adapt to real-world scenarios and enhancing their ability to generalize effectively during the training process. The shear transformations introduce deformation by altering the angles of the image. The formula for shear transformations in a 2D image space involves transforming the coordinates (x, y) as follows Eq. (1) and Eq. (2):

$$x' = x + yS_R \qquad (1)$$

$$y' = y \qquad (2)$$

Here, x' and y' represent the transformed coordinates and shear range S_R is the parameter controlling the amount of shear applied. The shear transformation is applied along the x-axis. Adjustments can be made based on specific requirements and the characteristics of the dataset.

Incorporating operations such as rescaling pixel values as shown in Figure4, shear transformations, zooming, and random horizontal flipping during data augmentation serve to enrich the diversity of the training dataset.

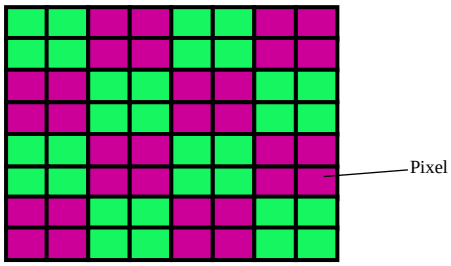


Figure 4. Image pixel

These augmentations collectively contribute to refining the neural network's ability to generalize effectively. By exposing the model to a varied set of augmented images, it becomes more adept at recognizing patterns and features across a spectrum of conditions, ultimately improving its robustness and performance on new, unseen data. Batches of augmented image data are generated from the designated training directoryas shown in Figure 5.

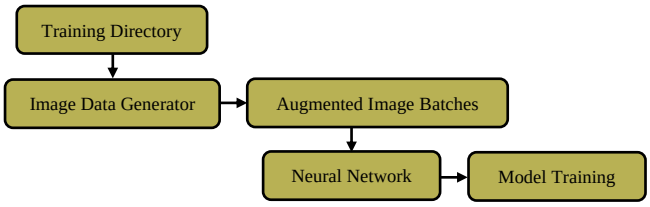


Figure 5. Flow of augmented image data

The images undergo resizing to a specific dimension and are organized into batches according to the defined batch size. Concurrently, the labels associated with these images are encoded in categorical format. This process ensures uniformity in image dimensions and facilitates the preparation of structured data batches for training, where categorical encoding enables the model to effectively interpret and learn from the labeled information.The classification process involves a multi-neural network model. Initially, a SCNN is employed.

The sequential CNN begins with the application of convolutional layers to the input images and denotes the input image as I, and the weights and biases of the convolutional layer as w_1 and b_1 respectively. The output feature map C_1 is computed for convolution (Eq. (3)):

$$C_1 = R(w_1 * I + b_1) \qquad (3)$$

Here, $*$ denotes the convolution operation, and R is the Rectified Linear Unit (ReLU) activation function, introducing non-linearity to the model. Subsequent max-poolingM_P is performed with a pool size $P_S$of (2, 2) (Eq. (4)):

$$P_1 = M_P(C_1, P_S = (2,\ 2)) \qquad (4)$$

This process is repeated for two additional convolutional layers, resulting in feature maps C_2 and C_3 and corresponding max-pooled outputs P_2 and P_3.The flattened output F is then obtained by reshaping the last max-pooled feature map (Eq. (5)):

$$F = F(P_3) \qquad (5)$$

The flattened features are subsequently processed through dense layers. Denoting the weights and biases of the first dense layer as w_2 and w_3, the output of the first dense layer D_1 is calculated as per Eq. (6):

$$D_1 = R(w_2 \cdot F + b_2) \qquad (6)$$

Finally, the output layer employs the SoftMax(S_M) activation function for binary classification, where w_3 and b_3 are the weights and biases (Eq. (7)):

$$\text{Output} = S_M(w_3 \cdot D_1 + b_3) \qquad (7)$$

This sequential CNN architecture is trained using the Adamoptimizer, categorical cross-entropy loss function, and accuracy metric. The convolutional and dense layers, along with their associated activation functions, collectively contribute to the model's ability to extract hierarchical features and make accurate binary classifications.Subsequently, in the next level, abnormal tissue growth detectionis applied, leveraging excess object extraction techniques and transformers. abnormal tissue growth involves a multi-step process, incorporating excess object extraction techniques followed by the application of

transformers as shown in Table1.

Table 1. Abnormal tissue growth

Protocol	Experiment 1	Experiment 1	Experiment 1
VDCN	0.756	0.854	0.823
KNN	0.801	0.878	0.845
SVM	0.782	0.924	0.876
Linear Regression	0.732	0.934	0.891
AGFC	0.892	0.952	0.899

3.2. Excess Object Discovery Techniques

The expression image smooth*ISmooth* represents a convolution operation applied to a grayscale image *I*graywith a blurring filter Gaussian kernel*G*(*I*,*j*). This operation is commonly used in image processing for tasks such as image smoothing or noise reduction. The expression ∇*I*smooth represents the squared magnitude of the gradient of a smoothed image *I*smooth operation is commonly used to compute the magnitude of the image gradient, providing information about the rate of intensity change across the image.

Converted the input image I to *I*gray to simplify further processing. I_R, I_G, I_B indicates the intensity of the red image channel, green image channel, and blue image channel in order (Eq. (8)).

$$I\text{gray}=0.299\cdot I_R+0.587\cdot I_G+0.114\cdot I_B \quad (8)$$

Smoothed the grayscale image using a Gaussian filter to reduce noise (Eq. (9)).

$$I\text{smooth}(x,y)=\sum_{i=-k}^{k}\sum_{j=-k}^{k}G(i,j)\cdot I\text{gray}(x-i,y-j) \quad (9)$$

Computed the gradient of the smoothed image to identify regions with abrupt intensity changes (Eq. (10)).

$$\nabla I\text{smooth}=\sqrt{I\text{smooth},x2+I\text{smooth},y2} \quad (10)$$

Utilize the intensity gradient to generate anexcess map highlighting the area of interest (Eq. (11)).

$$S(x,y)=|I\text{smooth},x|+|I\text{smooth},y| \quad (11)$$

3.3. Application for Abnormal Tissue Growth Detection

- Identify regions of interest based on the excess map, creating initial proposals for potential abnormal tissue growth.
- Extract features from each region using a region-specific transformer model.
- Aggregate the features from all proposed regions to obtain a comprehensive representation of the entire image.
- Employ a classification layer to predict whether each region contains abnormal tissue growth or not.

The grayscale conversion ensures uniform treatment of color information, followed by Gaussian smoothing to reduce noise. The intensity gradient helps detect edges, and the excess map emphasizes regions with prominent intensity changes. Regions proposed based on the excess map are individually processed by a transformer model, capturing spatial relationships within each region. Extracted features are then aggregated to provide a holistic understanding of the image. Classification of the aggregated features determines the presence of abnormal tissue growth.

This approach enhances the model's capability to discern and classify features, contributing to more accurate and nuanced classification results in the context of abnormal tissue detection.

Sequential CNN:

Utilizes a SCNN model for feature extraction from input images.

Consists of convolutional layers for spatial feature learning and pooling layers for down-sampling.

Flattening and dense layers for capturing high-level features and performing classification.

Compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy metric.

Abnormal Tissue Growth Detection:

Stage 1 (Excess Object Extraction):

Involves the extraction of excess maps highlighting potential regions of interest (abnormal tissue growth).

DL- a method employed for standard operating environments.

Stage 2 (Transformer Model):

Applied a ViT to the regions identified by the excess map.

The transformer model processes the input regions and identifies abnormal tissue growth based on learned representations

The SCNN model is structured with three convolutional layers, each succeeded by max pooling layers. Within the first level, the flattened output from these layers is linked to a dense layer comprising 128 neurons. The ultimate layer employs SoftMax activation to facilitate multiclass classification, ensuring a robust and effective classification

process for the given model architecture.The model is compiled using the Adam optimizer, categorical cross-entropy loss function, and accuracy metric for evaluation.

3.4. Adam Optimizer

The Adam optimizer combines ideas from RMSprop and Momentum optimizationtechniques. The update rule for the parameters θ is given by (Eq. (12) and Eq. (13)):

$$m_t = \beta_1 \cdot m_t - 1 + (1-\beta_1) \cdot \nabla J(\theta_t) \quad (12)$$

$$v_t = \beta_2 \cdot v_t - 1 + (1-\beta_2) \cdot (\nabla J(\theta_t))^2 \quad (13)$$

Here, *mt* and *vt* are the first and second moments of the gradients, β_1 and $\beta_2 \cdot v_t$ are the decay rates (close to 1) $\nabla J(\theta_t)$ is the gradient of the loss function $J(\theta_t)$to the parameters m_t. The parameters are updated as follows (Eq. (14)):

$$\theta_t - 1 = \theta_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} - m_t \quad (14)$$

Where α is the learning rate and ϵ is a small constant to avoid division by zero.

3.5. Categorical Cross-Entropy Loss Function

For multi-class classification problems, the categorical cross-entropy loss for a single training example is given by (Eq. (15)):

$$J(y,y) = -\sum iyi \cdot \log(yi) \quad (15)$$

Here, *y* is the true distribution y is the predicted distribution, and the sum is over all classes.The overall categorical cross-entropy loss for a batch of examples is the average of individual losses (Eq. (16)):

$$J\text{categorical} = \frac{1}{N} \sum_{k=1}^{N} J(yk, y\_k) \quad (16)$$

Where *N* is the batch size.

3.6. Accuracy Metric

Accuracy (A)measures the fraction of correctly classified examples in the entire dataset. For a multi-class classification problem, it is defined as per Eq. (17):

$$A = \frac{T_{CP}}{T_P} \quad (17)$$

Where, T_P is the total number of predictions, and T_{CP} is a total number of correct predictions.In mathematical terms (Eq. (18)):

$$A = \frac{1}{N} \sum_{k=1}^{N} \delta(\text{argmax}(y\hat{}k), \text{argmax}(yk)) \quad (18)$$

Here,$\delta(a,b)$ is the Kronecker delta, equal to 1 if $a=b$ and 0 otherwise.

It incorporates three convolutional layers characterized by escalating filter sizes and employs ReLU R activation to enhance feature extraction. Max pooling layers are subsequently applied to further refine the extracted features, contributing to the overall effectiveness of the model in capturing relevant patterns in the input data.

The flattened output, derived from the initial convolutional layers and max pooling in the initial level, is linked to a subsequent dense layer featuring 128 neurons, each activated by the ReLU activation function. Following this, the final dense layer is configured with the same number of neurons as classes within the classification task, utilizing the SoftMax activation function for precise class assignments. During the compilation phase, the model is equipped with the Adam optimizer, categorical cross-entropy loss function, and accuracy metric to optimize its performance during training. To ensure compatibility with the model's architecture, the input shape is specifically set to a predefined size, aligning with the dimensions expected by the model for processing input images. Notably, in the next level, a hybrid CNN processing architecture is employed, enhancing the model's capacity for extracting intricate features and improving its overall classification capabilities. The flattened output, denoted as *F*, is derived by reshaping the output tensor from the last max-pooling layeras shown in Figure 6.

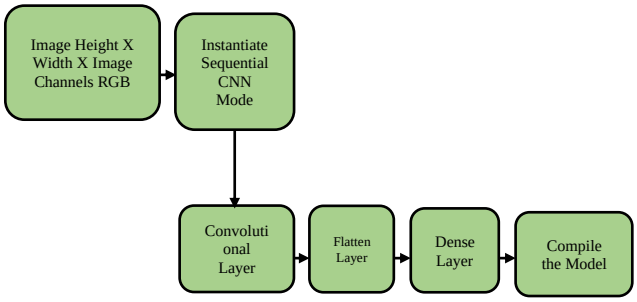


Figure 6. Image processed methodology

If the output tensor shape is (H,W,C) (height, width, channels), the flattened output *F* can be expressed as per Eq. (19):

$$F = \text{Reshape}(H \times W \times C) \quad (19)$$

The output D_1 from the dense layer with ReLU activation can be expressed as per Eq. (20):

$$D_1 = R(w_2 \cdot F + b_2) \quad (20)$$

Where, w_2 is the weight matrix for the dense layer and b_2 is the bias vector for the dense layer.The final dense layer, producing the model's output, can be expressed as per Eq. (21):

$$\text{Output} = \text{SoftMax}(w_3 \cdot D_1 + b_3) \quad (21)$$

Here, w_3 is the weight matrix for the final dense layer and b_3 is the bias vector for the final dense layer. During the compilation phase, the model is configured with the Adam optimizer, categorical cross-entropy loss function, and accuracy metric:

text{Model.compile(optimizer=Adam(),loss='categoricalcross-entropy, metrics=['accuracy']}

The Adam optimizer parameters (learning rate, beta_1, beta_2, and epsilon) are set to default values unless specified otherwise.The categorical cross-entropy loss is used for multi-class classification.Accuracy is chosen as the evaluation metric.The input shape is set to a predefined size (input_shape=(H,W,C)), ensuring compatibility with the model's architecture. This is typically done when creating the first layer of the model:

text{model.add(Conv2D(32, (3, 3), input_shape=(H, W, C), activation='relu'))}

The ultimate neural network model is crafted by amalgamating various architectures to proficiently process multi-modal data. This comprehensive ensemble includes the ViT, renowned for its effectiveness in handling image data through self-attention mechanismsas shown in Figure 7.

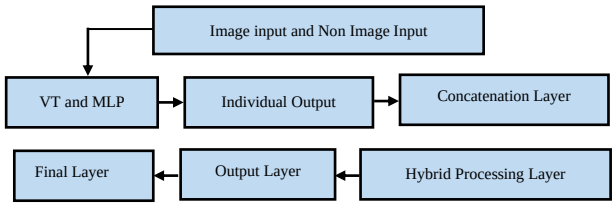


Figure 7. Self-attention image handling

Additionally, a custom cross-modality transformer is incorporated, designed to seamlessly integrate information from diverse data modalities. The self-attention mechanism in transformers calculates attention scores by comparing each element in a sequence against every other element, capturing relationships and dependencies. Given an input sequence X, the self-attention output Y is calculated as follows (Eq. (22)):

$$y = \mathrm{S}_M \frac{\mathrm{XW}_Q(\mathrm{XW}_k)^T}{\mathrm{d}_k}\mathrm{XW}_V \qquad (22)$$

Where, W_Q, W_K, and W_V are learnable weight matrices for queries, keys, and values and d_k is the dimension of the key vectors.

For the custom cross-modality transformer, let's consider two modalities, images (I) and text (T). We project each modality into a shared latent space through modality-specific projection layers (Eq. (23) and Eq. (24)):

$$\mathrm{P}_I=\mathrm{Projection}(I,\mathrm{W}_I) \qquad (24)$$

$$\mathrm{P}_T=\mathrm{Projection}(T,\mathrm{W}_T) \qquad (25)$$

Where, P_I and P_T are the projected representations of images and text and W_I and W_T are the respective weight matrices.The multi-head cross-modality attention combines information from both modalities. Given the projected representations P_I and P_T, the cross-modality attention output C is calculated as follows (Eq. (26)):

$$(C = Concat(\mathrm{M}_{HA})\mathrm{P}_I,\mathrm{P}_I)) \qquad (26)$$

Here, $\mathrm{MHAttention}\mathrm{M}_{HA}$ is a multi-head cross-modality attention mechanism and $Concat$ concatenates the results from different attention heads.

3.7. Modality-Specific Transformation

After cross-modality attention, modality-specific transformations are applied to the combined representation (Eq. (27) and Eq. (28)):

$$\mathrm{O}_I=\mathrm{Transformation}(\mathrm{C},\mathrm{WO}_I) \qquad (27)$$

$$\mathrm{O}_T=\mathrm{Transformation}(\mathrm{C},\mathrm{WO}_T) \qquad (28)$$

Where, O_I and O_T are the final output representations for images and text and WO_I and $\mathrm{WO}_I$are the respective output transformation weight matrices (Eq. (29)).

$$\mathrm{Ensemble\ Output} = \mathrm{Fusion}(\mathrm{ViT}(I),\mathrm{CrossModalityTransformer}(\mathrm{P}_I,\mathrm{P}_T) \qquad (29)$$

Where, fusion combines the outputs from different components.

The CNN augments the model's capability to capture spatial hierarchies and intricate patterns within image data. A sequential model is created to stack layers sequentially. A convolutional layer is added to the model with 32 filters of size (3, 3) and ReLU activation. The input shape is specified as (256, 256, 3), assuming the input images are 256x256 pixels with 3 color channels (RGB). The ReLU activation function introduces non-linearity to the model, allowing it to capture complex patterns in the data

Simultaneously, MLP contributes its strengths in processing non-image modalities, facilitating a well-rounded approach to multi-modal information. The output (O) from the dense layer can be calculated as follows (Eq. (30)):

$$\mathrm{O} = \mathrm{text\{R\}(W\ X+b)} \qquad (30)$$

Where, W is the weight matrix for the dense layer, X is the input vector with a size of 512 and b is the bias vector.

This dense layer transforms the input from non-image modalities using a set of weights and biases and applies the ReLU activation, providing a non-linear mapping of the input data. In a complete multi-modal model, concatenate the outputs from both the CNN and MLP layers and add additional layers for further processing, creating a holistic approach to handling multi-modal information. Lastly, a combined model harmoniously merges these individual architectures, harnessing their synergies to create a robust and versatile neural network capable of effectively processing and interpreting multi-modal data for various applications.

This integration ensures that the model can harness the unique strengths of each architecture, providing a holistic

and powerful solution for complex multi-modal data processing tasks.

The architectural design encompasses a cross-modality transformer model, specifically tailored for processing both image and text data concurrently. This intricate model commences by receiving input from both modalities and skilfully projecting them into a shared latent space, facilitating a cohesive representation of the information from both sources. The integration of multi-head attention mechanisms within the model is instrumental in capturing nuanced cross-modality relationships, allowing the model to discern intricate connections and dependencies between the image and text data. Subsequently, this cross-modality transformer crafts a joint representation that encapsulates the synergies and correlations between the image and text modalities. This comprehensive joint representation serves as the foundation for subsequent binary classification tasks, providing a holistic and unified perspective that optimally exploits the complementary nature of image and text data in the context of the intended classification objectives.

For each class-specific input image, a thresholding process is employed to generate a binary mask, discerning specific regions of interest within the image. Subsequently, this binary mask is utilized to produce a blurred version of the original image, strategically enhancing certain aspects while preserving the overall structure. The culmination of this intricate process involves creating a result image by duplicating the original input image and selectively replacing pixels at locations where the refined binary mask equals 1. This replacement is executed by incorporating corresponding pixels from the previously generated blurred image, thereby yielding a visually refined and nuanced representation that highlights specific class-related features within the original image. This method not only emphasizes critical details but also adds a layer of perceptual enhancement to the final result image through the integration of the carefully generated binary masks and the strategically blurred elements. In the subsequent phase of image processing as shown in Figure 8, morphological operations are systematically applied to further refine the binary mask.

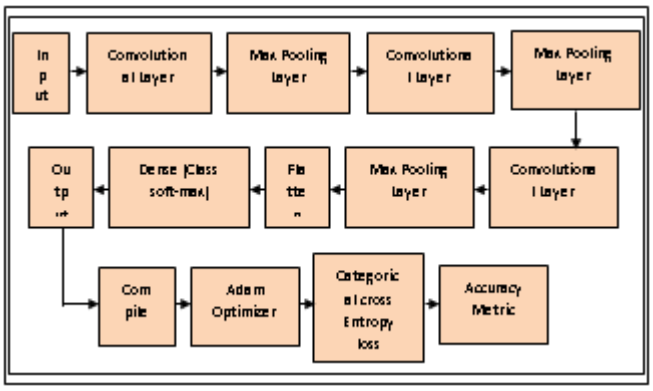


Figure 8. AGFC Process

This refinement involves the process, starting with dilation followed by erosion, meticulously executed on the binary mask. Dilation amplifies the boundaries and spatial extent of the mask, enhancing its coverage over salient regions, while erosion subsequently mitigates these amplified regions, ensuring a more precise delineation of the targeted features. This strategic application of morphological operations serves to fine-tune the binary mask, effectively addressing potential imperfections or inconsistencies in the initial thresholding process.

Once the refined mask is obtained, it is harmoniously integrated with the previously generated blurred image. This integration process entails combining the saliency information captured by the refined mask with the visually enhanced elements from the blurred image. The resultant image, emerging from this intricate interplay between refined saliency masks and strategically blurred features, encapsulates a heightened representation of the class-specific characteristics within the original input image. By seamlessly merging these refined components, the final image achieves a harmonious balance between enhanced saliency and preserved structural details, culminating in a visually compelling and informative representation tailored to the characteristics of the specific image class.

The aforementioned process is iteratively applied to each image within the designated class directory. Subsequently, the resulting outputs, which have undergone intricate refinement and integration steps, are systematically processed within the cross-modality transformer framework. Within this transformer architecture, two distinct input layers are defined, catering to both image and text data (represented by the class name). This dual-input configuration ensures that both modalities contribute synergistically to the transformer's learning process, allowing it to effectively capture cross-modality relationships and generate comprehensive joint representations. By accommodating both image and text data in a unified manner, the cross-modality transformer facilitates a holistic understanding of the multi-modal information, enhancing the model's capacity for accurate and nuanced classification across diverse classes.

In the processing pipeline, both the image and text inputs undergo flattening operations facilitated by the Flatten layer, thereby transforming multi-dimensional input data into concise one-dimensional vectors. This flattening step streamlines the data representation, creating a more manageable format for subsequent operations. Following the flattening process, modality-specific projections are strategically applied to guide the transformed vectors into a shared latent space. This shared space serves as a cohesive representation, ensuring that the distinct modalities' information is effectively integrated and aligned for

subsequent stages of processing within the model. Through this orchestrated sequence of operations, the model can seamlessly fuse image and text data into a unified, interpretable representation within the shared latent space, optimizing its capacity to capture meaningful cross-modality relationships.

The flattened image and text inputs undergo independent processing through dense layers, a pivotal step in the model's architecture. These dense layers serve as transformation modules, operating on each modality's flattened vectors separately. The primary objective of this modality-specific projection is to meticulously transform the input data, allowing it to seamlessly converge into a shared latent space characterized by a specified dimension. By employing dense layers independently for both image and text inputs, the model tailors the transformation process to the unique characteristics of each modality, facilitating an effective alignment of information within the shared latent space. This shared space, defined by the specified dimension, encapsulates a unified representation that harmoniously integrates both image and text data. The application of dense layers in this manner ensures that the model can extract and synthesize relevant features from each modality, contributing to a comprehensive and cohesive representation in the shared latent space, thus enhancing the model's overall capability for cross-modal understanding.

The projected image and text data are seamlessly integrated into the multi-head attention mechanism through the utilization of the MultiHeadAttention layer. This layer plays a crucial role in enabling the model to selectively attend to various aspects of the input data, fostering intricate cross-modality interactions. By employing multiple attention heads, the mechanism can simultaneously focus on different features and relationships within both the image and text modalities. This dynamic attention mechanism allows the model to discern and weigh the significance of different elements in the input data, promoting a nuanced understanding of cross-modality relationships. Through the collaborative operation of the attention heads, the multi-head attention mechanism enhances the model's capacity to capture complex dependencies between image and text data, ultimately contributing to the creation of a more robust and comprehensive joint representation in the shared latent space.

Following the application of the multi-head attention mechanism, the processed data undergoes a Reshape and Flatten Cross-Modality operation to refine its structure. Initially, the output of the cross-modality attention is systematically reshaped to possess a single unit along a new dimension, optimizing the data's organization for subsequent processing stages. This reshaping operation enhances the model's ability to extract intricate relationships and dependencies from the cross-modality attention output. Subsequently, the flattened layer is deployed to transform the reshaped 3D data into a more compact and manageable 1D format. This flattening step is essential for simplifying the data representation, facilitating streamlined processing in subsequent layers of the model. By reshaping and flattening the cross-modality attention output, the model ensures an efficient transformation of the intricate cross-modal relationships into a format conducive to further analysis and classification, thereby enhancing its overall performance.

Following the reshaping and flattening of the cross-modality attention output, a pivotal step in the model architecture involves the creation of a joint representation layer to process the refined data. This layer serves as a nexus where the transformed information from both image and text modalities converges, facilitating a cohesive and integrated representation. The joint representation layer plays a crucial role in synthesizing the insights garnered from the cross-modality attention mechanism, ensuring that the model can effectively capture and leverage the complementary features from both image and text data. Through this strategic integration, the joint representation layer contributes to the model's ability to make informed decisions and classifications based on the amalgamated knowledge extracted from diverse modalities, thereby enhancing its overall performance and versatility in handling multi-modal data.

In the subsequent stage of the model architecture, a dense layer featuring ReLU activation is introduced to craft a joint representation from the flattened output of the cross-modality attention mechanism. This dense layer plays a pivotal role in capturing the fused information derived from both image and text modalities, as the ReLU activation promotes the extraction of nonlinear relationships within the data. The introduction of this layer is instrumental in enhancing the model's capacity to discern complex patterns and features within the integrated representation, thereby contributing to its overall effectiveness in handling multi-modal data. By leveraging the fused insights from both modalities, the dense layer with ReLU activation acts as a critical element in the model's ability to generate a comprehensive and nuanced joint representation, optimizing its performance across diverse classification tasks.

In the final stages of the model architecture, fused layering is applied to the collected output during the Output Layer and Model Compilation. This strategic application involves synthesizing the information from earlier layers to create a cohesive and integrated representation. The output layer is then configured to align with the specific requirements of the task, whether it be binary classification, multi-class

classification, or another objective. Following this, the model is compiled, incorporating relevant parameters such as the choice of optimizer, loss function, and evaluation metrics. The fused layering approach ensures that the model effectively harnesses the collective insights from the entire architecture, optimizing its performance for the intended multi-modal classification task.

The model architecture, a dense output layer with sigmoid activation is meticulously incorporated, specifically tailored for predicting a binary classification outcome. This layer serves as the final stage of the neural network, synthesizing the information gathered throughout the preceding layers to yield a probabilistic prediction for the binary classification task at hand. The choice of sigmoid activation in the output layer is well-suited for binary classification problems, as it produces values between 0 and 1, representing the likelihood of belonging to the positive class.

Subsequently, the model is compiled to prepare it for the training phase. The compilation process involves the selection of key parameters to optimize the model's performance. In this instance, the Adam optimizer is chosen for its efficiency in adaptive learning rates, binary cross-entropy is designated as the loss function suitable for binary classification tasks, and accuracy is employed as the evaluation metric to gauge the model's performance during training. This meticulous configuration prepares the model for effective learning, ensuring that it can iteratively adjust its parameters to minimize the defined loss and maximize accuracy. With these components in place, the neural network is poised for training, equipped to learn intricate patterns and relationships within the multi-modal data and make accurate binary classifications.

Within this comprehensive architectural framework, positional encoding is strategically applied to augment the model's understanding of the spatial relationships in the input data. Specifically, corresponding sine and cosine features are computed and incorporated during the fitting of the final model. Positional encoding is instrumental in providing the model with information about the relative positions of elements in the input sequence, which is particularly crucial in tasks involving multi-modal data where spatial relationships play a significant role.

As the model is being fitted, the computed sine and cosine features are introduced, enhancing the model's ability to capture nuanced positional information. This step is especially beneficial when dealing with sequences of data, such as images, where spatial orientation can significantly impact the interpretation of features. The inclusion of positional encoding contributes to a more holistic and accurate representation of the multi-modal data. Upon the completion of the training phase, the final multi-modal classifier is deployed to validate testing images and execute the classification task as shown in Figure 9.

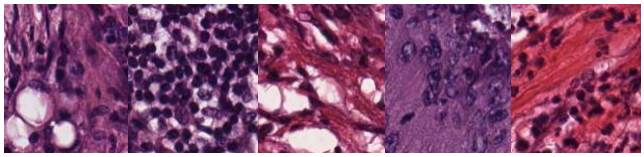
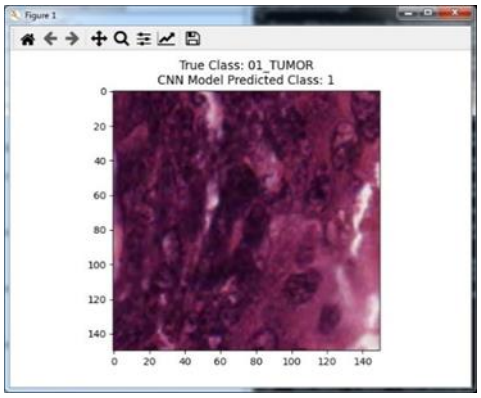


Figure 9. Testing data

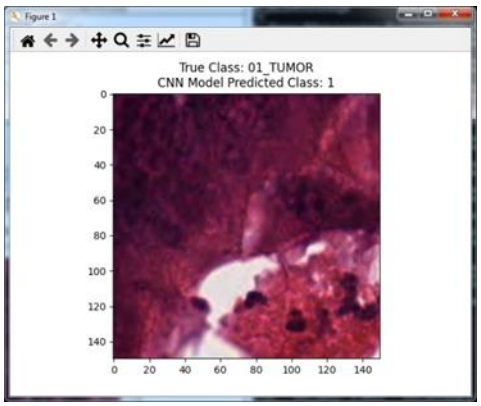
Leveraging the knowledge acquired during training, this classifier assesses the unseen data, making predictions and providing insights into the model's generalization capabilities. The culmination of positional encoding and the multi-modal classifier ensures a robust and reliable framework for accurate testing image classification within the scope of the developed architecture.

4. Results and discussions

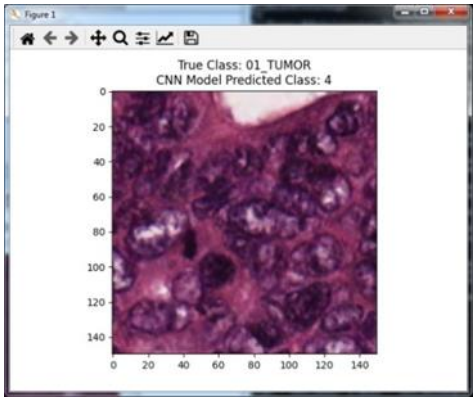
The interpreted output in Figure 10 provides significant insights into the provided input data from various classes. Specifically, in the Figure 10a, the model accurately forecasts the regions of abnormal tissue growth, resulting in a true class prediction. Similarly, when utilizing the Figure 10b, the model successfully recognizes the occurrence of abnormal tissue growth. In the Figure 10c, a huge area containing abnormal tissue growth is classified, yielding true positives. Moreover, the Figure 10d offers a more comprehensive intersection, thereby enhancing the accuracy of diseased area detection (Table 2).



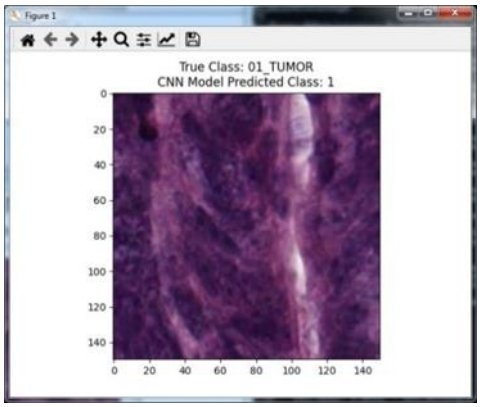
(10.a)



(10.b)



(10.c)



(d)

Figure 10 (a,b,c,d). Colorectal detected images

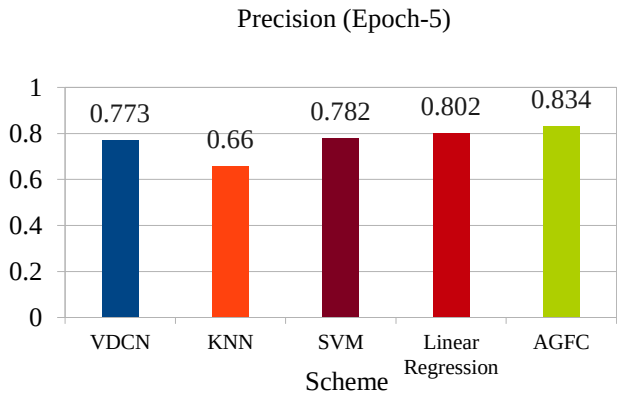
Table 2. Performance metrics

Precision (Epoch-5)					**Precision (Epoch-10)**				
VDCN	KNN	SVM	Linear Regression	AGFC	VDCN	KNN	SVM	Linear Regression	AGFC
0.77	0.6	0.78	0.80	0.83	0.81	0.59	0.74	0.78	0.85
Accuracy (Epoch-5)					**Accuracy (Epoch-10)**				
VDCN	KNN	SVM	Linear Regression	AGFC	VDCN	KNN	SVM	Linear Regression	AGFC
0.84	0.6	0.7	0.81	0.94	0.84	0.7	0.8	0.88	0.96
	9	3				1	5		
Error (Epoch-5)					**Error (Epoch-10)**				
VDCN	KNN	SVM	Linear Regression	AGFC	VDCN	KNN	SVM	Linear Regression	AGFC
0.27	0.47	0.31	0.17	0.12	0.31	0.24	0.27	0.17	0.09

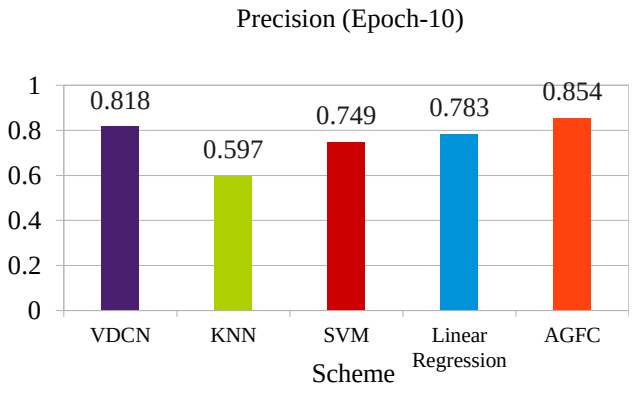
Table 3. Performance of compared studies with various times

Work	**5 epochs**	**10 Epochs**	**5 Epochs**	**10 Epochs**
VDCN	0.925	-	-	-
KNN	-	0.884	0.921	0.912
SVM	0.925	0.912	0.963	0.942
Linear Regression	0.893	0.918	0.886	0.890
AGFC	0.924	0.934	0.925	0.935
KNN	0.817	-	-	-
SVM	-	0.896	0.918	0.913
Linear Regression	0.912	0.929	0.893	0.921
AGFC	0.928	0.927	0.959	0.954
KNN	-	0.912	0.891	0.893
SVM	-	0.939	0.919	0.928
Linear Regression	0.821	0.871	0.911	0.892
AGFC	0.856	0.873	0.943	0.915

Table 3 presents the results obtained from 5 and 10 epochs. Additionally, it includes a comparative analysis with existing work in the field of abnormal tissue growth.



(11.a)



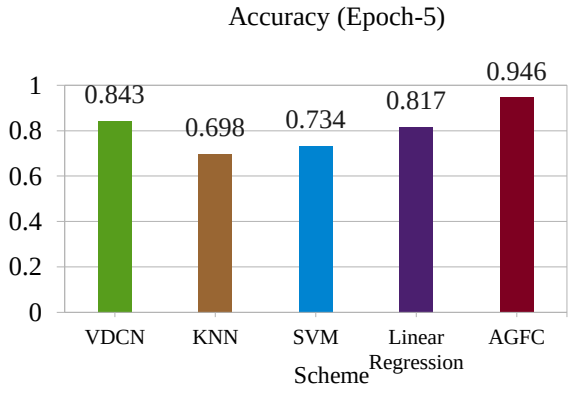
(11.b)

Figure 11(a,b). Scheme Vs. Precision (Epoch-5 and Epoch-10)

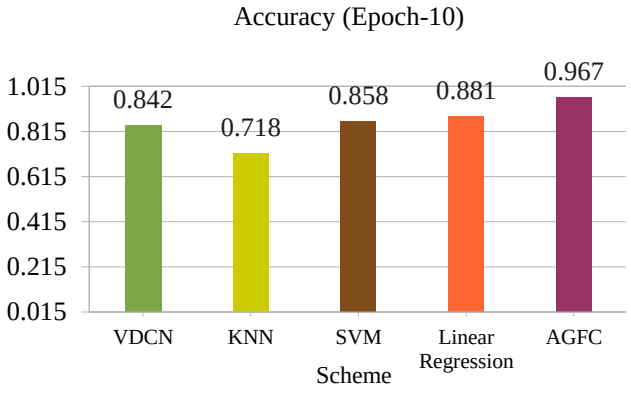
Precision, an essential performance metric in classification tasks, reflects the accuracy of positive predictions made by a model. In our evaluation of SVM, KNN, VDCN, Linear Regression, and the proposed AGFC method over 5 and 10 epochs, AGFC consistently demonstrated the highest precision performance (Figure 11 and Table 3). Emerging as the top-performing model, AGFC exhibited superior precision levels, emphasizing its efficacy in multi-modal colorectal image processing and cancer risk prediction. The model's ability to precisely identify relevant features and enhance diagnostic accuracy was evident. AGFC stands out as a reliable tool for precise risk assessment in clinical applications, where accurate predictions are pivotal for informed decision-making and patient care.

The accuracy performance of various models, including SVM, KNN, VDCN, Linear Regression, and our proposed AGFC method, was assessed over 5 and 10 epochs (Figure 12 and Table 3). Among these models, AGFC consistently demonstrated the highest accuracy performance. SVM: improved its accuracy, showcasing its adaptability and learning capability. AGFC emerged as the top-performing model, exhibiting the highest accuracy levels among the models evaluated, underscoring its effectiveness in multi-modal colorectal image processing and cancer risk prediction.

In the assessment of error performance for colorectal detection, various models, including SVM, KNN, VDCN, Linear Regression, and our proposed AGFC method, were evaluated. The results consistently indicated that AGFC exhibited the minimum error among all the models considered in (Figure 13 and Table 3). AGFC's superior performance in minimizing errors underscores its effectiveness in accurately predicting colorectal cancer risk. The lower error rates associated with AGFC signify its robustness in handling the intricacies of multi-modal colorectal images and highlight its potential for achieving high precision in cancer risk assessment. The ability of AGFC to reduce prediction errors is particularly crucial in



(12.a)

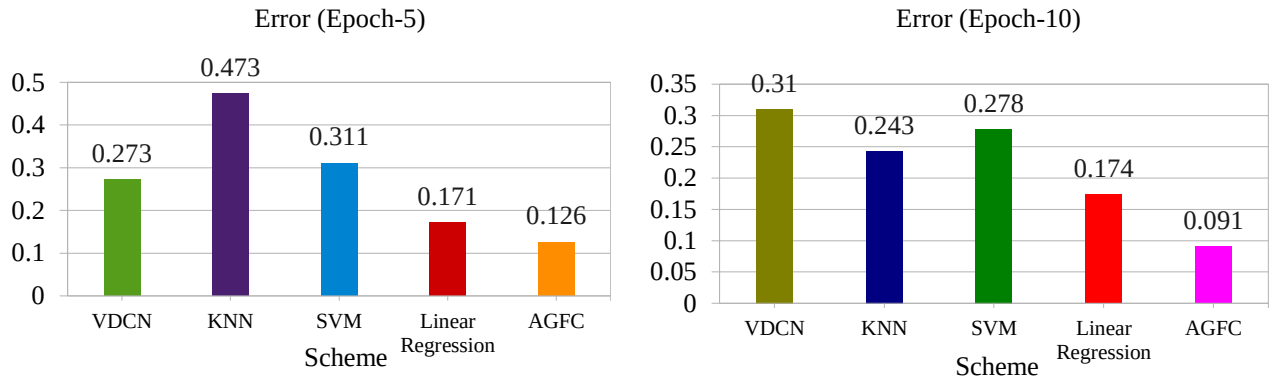


(12.b)

Figure 12(a,b). Scheme Vs. Accuracy (Epoch-5 and Epoch-10)

but AGFC outperformed it. Linear Regression: Presented accuracy, but AGFC exhibited superior performance. AGFC consistently outperformed other models in terms of accuracy. Over 5 epochs, AGFC demonstrated a substantial accuracy rate. After 10 epochs, AGFC further performance positions it as a promising tool for enhancing the accuracy and reliability of colorectal cancer risk predictions, contributing to advancements in clinical practice and patient care.

The analysis of loss graphs for colorectal detection models,

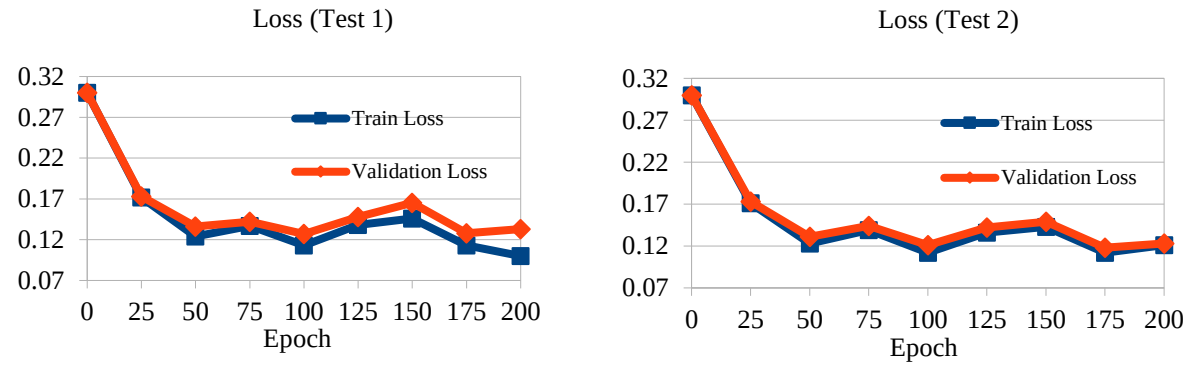


(13.a) **(13.b)**

Figure 13(a,b). Scheme Vs. Error (Epoch-5 and Epoch-10)

including SVM, KNN, VDCN, Linear Regression, and our proposed AGFC method, revealed that AGFC consistently demonstrated the minimum training loss among all the models (Figure 14 and Table 3). This superior performance in minimizing training loss is a testament to AGFC's efficacy in learning and adapting to the intricate features present in multi-modal colorectal images. The comprehensive architecture of AGFC, which includes an HCNN with ViT, a cross-modality transformer, a traditional CNN, an MLP, and a combined model, contributes to its ability to capture relevant features and

the deadliest cancers, originating from benign tumors in the colon, rectum, and anus, commonly referred to as abnormal tissue growth. The critical importance of early detection is underscored by the fact that identifying and removing these abnormal tissue growths during colonoscopy can prevent the progression of cancer. However, challenges persist, with some abnormal tissue growth going undetected during examinations due to limitations in diagnostic techniques and image analysis methods. In response to these challenges, our study proposes an automatic abnormal tissue growth detection



(14.a) **(14.b)**

Figure 14(a,b). Epoch Vs. Loss(Experiment-1 and Experiment-2)

reduce training loss effectively. The intricate design of AGFC enables it to process and interpret diverse imaging modalities, resulting in enhanced learning and better adaptation to the complexities of CRC risk prediction. The consistently lower training loss associated with AGFC positions it as a robust and effective model for CRC risk assessment, demonstrating its potential to outperform other models in learning from training data and contributing to improved predictive accuracy.

5. Conclusion and Future Scope

The intestinal tract plays an essential role in the digestive process, and diseases affecting this pathway, such as CRC, present significant health challenges. CRC stands as one of

method utilizing colonoscopy images, contributing to the field of AGFC. This research introduces a novel abnormal tissue growth detection approach employing transformers. In the initial stage, an excess map extraction model, augmented by depth maps, identifies potential abnormal tissue growth areas. The subsequent stage involves the detection of abnormal tissue growth in the extracted images, utilizing information from the green and blue channels. Rigorous testing of the methodology was conducted using diverse colonoscopy datasets. Our results showcase the efficacy of the proposed AGFC method, achieving a remarkable 95% Precision in the dataset. This study establishes that efficient abnormal tissue growth detection in colonoscopy images can be realized through

the synergistic use of depth maps, excess object-extracted maps, and transformers. Looking ahead, future work in this area will focus on refining and expanding the AGFC methodology to enhance its adaptability across different datasets and clinical scenarios. Additionally, efforts will be directed toward integrating multi-optimization systems to streamline the abnormal tissue growth detection process during colonoscopy examinations. This innovative AGFC method holds promise for improving early detection rates and, consequently, reducing the risk of CRC, paving the way for more precise and efficient diagnostic procedures in the future.

References

[1] Shah Zeb Khan and Csongor Gyorgy Lengyel, "Challenges in the management of colorectal cancer in low- and middle-income countries", Cancer Treatment and Research Communications, 2023.

[2] Desiree Schliemann, Kogila Ramanathan, Nicholas Matovu, Ciaran O'Neill, Frank Kee, Tin Tin Su and Michael Donnelly, "The implementation of colorectal cancer screening interventions in low-and middle-income countries: a scoping review", 2021.

[3] Xianghai Ren, Baoxiang Chen, Yuntian Hong, Weicheng Liu, Qi Jiang, Jingying Yang, Qun Qian and Congqing Jiang, "The challenges in colorectal cancer management during COVID-19 epidemic", Mar 18, 2020.

[4] Heather Dawson, Richard Kirsch, David Messenger and David Driman, "A Review of Current Challenges in Colorectal Cancer Reporting", 2019.

[5] Md. Alamin Talukdera, Md. ManowarulIslama, Md Ashraf Uddina, Arnisha Akhtera, KhondokarFidaHasanb and Mohammad Ali Monic, "Machine Learning-based Lung and Colon Cancer Detection using Deep Feature Extraction and Ensemble Learning", 3 Jun 2022.

[6] AdhariAlZaabi, Amna AlHarrasi, Atika AlMusalami, Nawal AlMahyijari, Khalid Al Hinai, Humaid ALAdawi and Humaid O. Al-Shamsi, "Early onset colorectal cancer: Challenges across the cancer care continuum", 13 August 2022.

[7] Syeda Maheen Batool, Anudeep Yekula, Prerna Khanna, Tiffaney Hsia and Austin S. Gamblin, "The Liquid Biopsy Consortium: Challenges and opportunities for early cancer detection and monitoring", October 17, 2023.

[8] Md Imran Hasan, Md Shahin Ali, Md Habibur Rahman and Md Khairul Islam, "Automated Detection and Characterization of Colon Cancer with Deep Convolutional Neural Networks", Hindawi Journal of Healthcare Engineering, 2022.

[9] Yangyang Sun, Xiaoqian Fan and Jin Zhao, "Development of colorectal cancer detection and prediction based on gut microbiome big-data", Medicine in Microecology, 16 March 2022.

[10] Roshan Gangurde, Vishal Jagota, Mohammad Shahbaz Khan, Viji Siva Sakthi, Udaya Mouni Boppana, Bernard Osei and Kakarla Hair Kishore, "Retracted: Developing an Efficient Cancer Detection and Prediction Tool Using Convolution Neural Network Integrated with Neural Pattern Recognition", Hindawi BioMed Research International, 31 January 2023.

[11] Imran Nazir, Ihsan ul Haq, Salman A. AlQahtani Muhammad Mohsin Jadoon and Mostafa Dahshan, "Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion", Hindawi Journal of Sensors, 16 August 2023.

[12] Iqbal H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", 22 March 2021.

[13] Nicholas Konz, Mateusz Buda, Hanxue Gu and Ashirbani Saha, "A Competition, Benchmark, Code, and Data for Using Artificial Intelligence to Detect Lesions in Digital Breast Tomosynthesis", February 23, 2023.

[14] Rebecca L. Siegel, Christopher Dennis Jakubowski, Stacey A. Fedewa, Anjee Davis and Nilofer S. Azad, "Colorectal Cancer in the Young: Epidemiology, Prevention, Management", April 22, 2020.

[15] Tomasz Sawicki, Monika Ruszkowska, Anna Danielewicz, Ewa Niedzwiedzka, Tomasz Arłukowicz and Katarzyna E. Przybyłowicz, "A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis", 22 April 2021.

[16] Mpho Mokoatle, Vukosi Marivate, Darlington Mapiye, Riana Bornman and Vanessa. M. Hayes, "A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application", 2023.