Measurement: Sensors 31 (2024) 100976

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Colorectal cancer detection based on convolutional neural networks (CNN) and ranking algorithm

A. Karthikeyan [a,*], S. Jothilakshmi [a], S. Suthir [b]

[a] Department of Information Technology, Annamalai University, Chidambaram, 608002, India
[b] Amrita School of Computing, Amrita VishwaVidyapeetham, Chennai, 601103, India

ARTICLE INFO

Keywords:
Colorectal cancer
CNN
Ranking algorithm
Loss prediction
Machine learning

ABSTRACT

With the development of targeted therapies, many treatments are based on molecular studies, which require sampling tumor tissue from paraffin blocks for sequencing. An automated solution could potentially reduce the workload of pathologists by acting as a screening device and may reduce the subjectivity in diagnosis. In tissue-based diagnostics, most of the work still needs to be done manually by a pathologist using a microscope to examine stained slides. The foundation of such tasks is to accurately distinguish cancer/malignant cells from normal/benign cells. However, the determination of tumor content is poorly reproducible with significant variation. As the size of tumor regions can be very small, pathologists are often required to use high magnification for detecting tumor cells. This requirement significantly increases the workload for pathologists. As digital pathology datasets have become publicly available and have opened up the possibility of evaluating the feasibility of applying deep learning techniques to improving the efficiency and quality of histologic diagnosis. The model proposed in this work is an application to detect colorectal cancer based on the Convolutional Neural Network and Ranking algorithm. Based on the performance evaluation, it is found that the proposed model is yielding better results in diagnosis of Colorectal Cancer than the existing methods in terms of Recall, Precision and Accuracy.

1. Introduction

In the modern era, cancer is the most spreading complex disease. Identifying cancer without biopsy at an early stage is further imperative. Moreover, taking a biopsy is not good for health as well. In general, cancer has been caused by hereditary instability and accumulation of multiple molecular alterations. It is also caused by cellular genes abnormal activation that controls cell growth or cell mitosis. Colorectal Cancer (CRC) is a cancer from uncontrolled cell growth in the colon or rectum. Colorectal cancer is also known as colon cancer, bowel cancer or colorectal adenocarcinoma. The main negative aspect of cancer is delayed diagnosis and treatment. Due to this problem, cancer has overtaken heart disease as the leading cause of death for any age on. Therefore, early detection of cancer is important [1,2].

Colorectal cancer, or cancer of the large intestine, is the third most frequent cancer in the world, with an annual incidence of 1.2 million cases and a 50 % fatality rate, making it the fourth leading cause of cancer-related fatalities. Furthermore, with 1.77 million people living with large bowel cancer, it is the second most common cancer after prostate cancer. CRC has traditionally followed urbanization trends, with more urbanized countries bearing the brunt of the CRC burden. Both genetic make-up (‘nature’) and lifestyle (‘nurture’) have a role in colorectal carcinogenesis, according to etiological research.

India has one of the lowest rates of CRC, with age-adjusted incidence rates of 4.2 and 3.2 for males and femalesrespectively. Indian researchers have found a consistent increase in CRC incidence, as well as an extremely high incidence of beginning at a young age without any obvious family history of cancer. Previous studies from highly populated developing nations in Asia [3] and elsewhere found a similar increase in the early age of onset CRC, possibly due to fast changes in lifestyle brought about because of urbanization.

The human large intestine, also known as the colorectum, is the last portion of the gastrointestinal system and is responsible for reabsorbing water and some minerals from undigested food, also known as chime. The chime enters the large intestine in liquid form, which is then processed in the colorectum by absorption of water, active secretion of mucin, and evacuation of bacteria from the gut flora to solid feces (which can take up to 16 h).

* Corresponding author.
E-mail addresses: keyanmailme@gmail.com (A. Karthikeyan), jothi.sekar@gmail.com (S. Jothilakshmi), s_suthir@ch.amrita.edu (S. Suthir).

https://doi.org/10.1016/j.measen.2023.100976
Received 4 May 2023; Received in revised form 19 November 2023; Accepted 14 December 2023
Available online 19 December 2023

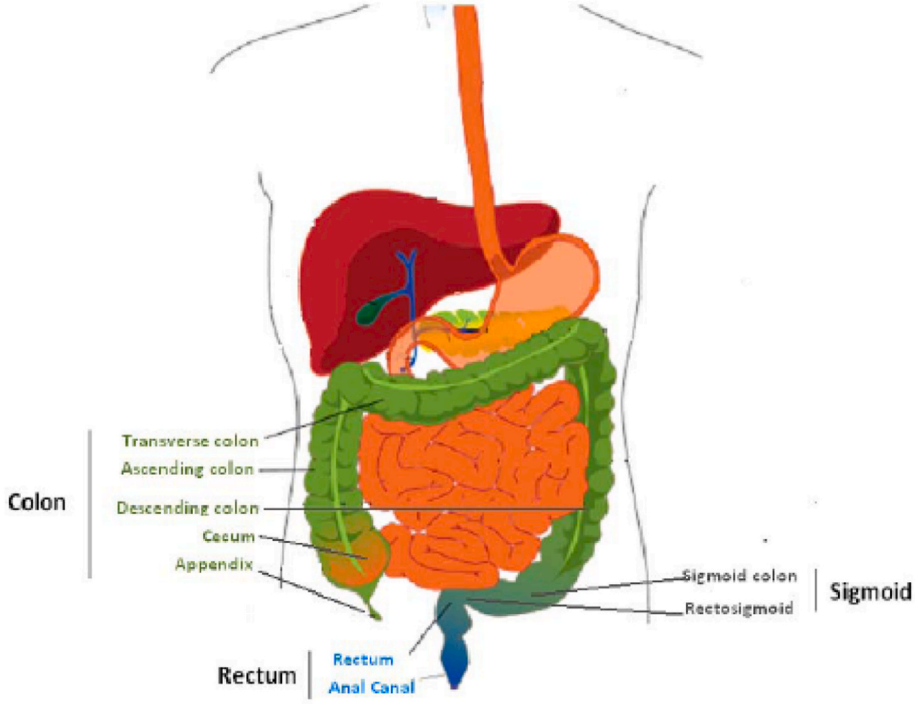


Fig. 1. Human digestive system and location of the large intestine. [SOURCE: 'Wikimedia commons'.].

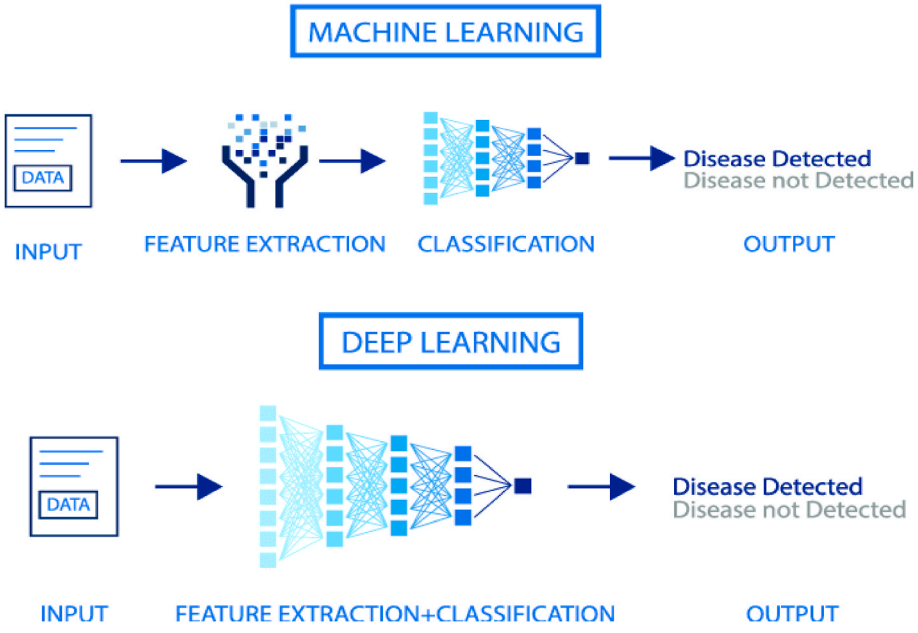


Fig. 2. Architecture of ML and DL models.

The human large intestine is around 1.5 m long and physically varies from the small intestine in that it is considerably larger in size but much shorter in length [4]. In addition to being shorter in length than the small intestine, the human large intestine lacks villi, which are little finger-like evaginations from the gut lining that help in the absorption of digested food. Furthermore, the big intestine has more goblet cells (mucin-secreting cells) than the small intestine [5]. The cecum, ascending colon (right colon), hepatic flexure, transverse colon, splenic flexure, descending colon (left colon), sigmoid colon, rectum, and anal canal are all parts of the large intestine as shown in Fig. 1.

Owing to the fact that the flow of blood from the intestine's wall and a considerablepiece of the rectum to the liver, colorectal cancer can metastasize to the liver after affecting the neighboring lymph nodes [6]. Of late, numerous computer-assisted diagnosis systems (CADs) were proposed for automated diagnosis of symptoms of cancer in the colon. The following Artificial Intelligence (AI) models are used for accurate diagnosis of cancer:

- Employing Machine Learning (ML) to recognize the colon cancer by analyzing thehistopathological images [7,8].
- Employing Deep Learning (DL) to recognize the colon cancer by analyzing thehistopathological images [9,10].

In MLmodels various stages such as preprocessing, feature extraction, feature selection and classification are used. Any noise that is included as a part of the image can be removed by the process of preprocessing. The preprocessed image is fed to the feature extraction module to extract the possible features from the data and the final classifier selects the most vital features from the set of features. The final decision is completedby means of a separate classifier in the classification stage. In DL based models, the feature extraction and classification stages are fused as a single stage in the deep model as represented in Fig. 2.

The proposed research work focuses on colorectal cancer detection using Convolutional Neural Network (CNN) and ranking algorithm. The manuscript is organized as follows:Section 2 provides a detailed survey about the existing techniques for the prediction of colorectal cancer. Section 3 describes the proposed methodology using CNN for the prediction of colorectal cancer. The results of the proposed method are compared with the existing models and the manuscript is concluded in the final section.

2. Literature review

A lot of research has been done for formulating techniques for the identification of various types of cancer at an early stage based on symptoms. Lung cancer is a lethal lung illness that affects the nodules of the lung. The proposed model consists of 20 parameters ordered based on the loss in weight, blood in mucus, back pain and other related factors [11]. Based on these factors, the model tries to diagnose the level of lung cancer and provides the result.

An ensemble framework was proposed to identify the possible occurrence of pneumonia from the chest X-ray Images [12]. The proposed ensemble model is a combination of Capsule Net and VGG-19 model to diagnose the pneumonia. Magnetic Resonance Imaging (MRI) images of colorectal cancer are used to detect the area in the model suggested by Jeroen B. Smaers et al. [13]. The exact stage of cancer is calculated based on the mean values of tumor area, and distance from tumor area to other parts. Theproposed model consists of pre-processing, clustering and segmentation of MRI images. By using the MRI based colorectal cancer identification model, the cancer was identified at the later stage rather than the initial stage.

Pallabi Sharma et al. [14] addressed the current problem of detecting the colorectal cancer in medical image processing from colonoscopy videos. As per the cancer statistics report, colorectal cancer is one of the most common cancers and the removal of precancerous cells from the large intestine is a crucial task. The manual process of diagnosis depends on the expertise of the medical practitioner. The authors proposed a two-stage classification model to detect colorectal cancer. The first stage extracts the features from the input images and the second stage performs the classification. The proposed model was compared with other existing CNN models namely VGG16, VGG19, Inception V3, Xception, GoogLeNet, ResNet50, ResNet100 and DenseNet. Based on the experimental results it is observed that the VGG19 model is the best deep learning method for colonoscopy image diagnosis.

NamitaSengar et al. [15] proposed an automated system for grading of colorectal cancer using image processing methods. The proposed model segments the glands automatically based on the threshold intensity rather than the manual model that requires expertise to analyze the histopathological tissues. Unlike most of the existing methods, the proposed model is fully automated and grades the images as benign healthy, benign adenomatous, moderately differentiated malignant and poorly differentiated malignant with an overall accuracy of 81 % on 165 histology images.

Tajbakhsh et al., 2015 [16] focused on Colonic polyp Detection and Classification using 3-way image production and CNN model. IndividualCNN based on features of the polyp candidate was applied on the candidate environs and the outcome was decided to either accept or reject the candidate. Bayramoglu et al. [17] depicted on Cell Nuclei Classification. The proposed method employed VGG-16 architecture. The layers clichéd from source networks are allowed to amend slowly whereas during the process of fine tuning, the features can learn at higher layers with elevated learning rates. The accuracy of prediction was 88.03 %.

Haj- Hassan et al. [18] employed the Multispectral Colorectal Cancer Tissue Classificationbased on active contour segmentation and CNNs. The model relies on CNNs for unsupervised learning butis unable to confinethe full inconsistency of tissues in the development of Colorectal cancerdue to the minimal size of the training data set. Hence the proposed model requires manual intervention in the process of diagnosis. Kainz et al. [19] suggested a work on Colon Gland Classification. The proposed method consists of two phases:Pre-processing is applied on the raw RGB image to normalize and extract the tissue composition; the second phase uses two trained classifiers namely Object-Net and Separator-Net. The former one is used for differentiating glandsfrom the background and the later one is used for unraveling the glands inthe input image.

Tomczak et al. [20] proposed an approach for colon cancer classification and achieved performance results in terms ofArea UnderCurve (AUC) similar to Gaussian process-based methods. The NOR operator employed in the model failed to gain improved accuracy and F1-score. Yoshida et al. [21] worked on colorectal biopsy histological based classification of specimens. The proposed model iteratively trained more than 1000 images to overcome the trade-off among the over fitting and under fitting problem. Certain request of data for validation was unsuccessful due to the feature distribution. The accuracy results in categorizing other diseases such as lymphoma or low-grade adenoma were not better than the existing models.

Bychkov et al. [22] portrayed the colorectal cancer classification based ontissue analysis which gave more precise outcomeforecast in contrast to the manual diagnosis model. In manual diagnosis model, the prognostic information available in the image is limited compared to the CAD model. One shortcoming of the proposed method is the time consumption for classification. Chen et al. [23] proposed a Deep Neural Network (DNN- CAD) for examining narrow-band images of minuscule colorectal polyps. The proposed model depends onextravagant narrow-band images of superior-quality classes alone. Thus, the results may be not as expected for the images with minimal quality or blur within the image.

Zhanget al. [24] characterized the research work on classification of colorectal cancer imagesalongside with additionalmedical diagnostic images. The proposed model is a combined deep and handcrafted visual feature (CDHVF) based algorithm that employs the features educated by a combined approach of two handcrafted descriptors i.e. Bag of Features (BoF) and Local Binary Pattern (LBP), and three pre-trained and fine tuned deep CNNs. The pre-trained Deep CNN models providedthe high dimensional deep features thattrounces by Principal Component Analysis.

Dabass et al. [25] focused on colon cancer classification with 31 layers of CNN architecture based on AlexNet. Accuracy for the two class classification model is less compared to other state-of-the artmodels. Manivannan et al. [26] concluded that the vital problem in cancer treatment is the early diagnosis of the disease. In general, cancer is diagnosed at the later stage rather than the initial stage. During the process of diagnosis, it would have compromised the vital organs of a human. The survey was concluded by establishing MRI as the majormeans for diagnosing and staging in patients with colorectal cancer.

Predicting colorectal cancer using machine learning algorithms encounters several challenges:

Data Quality and Quantity: Availability of high-quality data is essential for accurate predictions. In healthcare, obtaining comprehensive and diverse datasets with relevant features can be challenging due to privacy concerns, data silos, and limited sample sizes.

Imbalanced Datasets: Colorectal cancer datasets often have an imbalance between the number of cancerous cases and non-cancerous

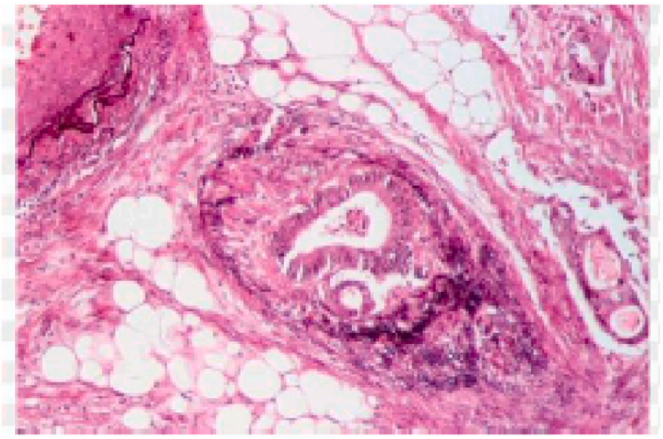


Fig. 3. Original image.

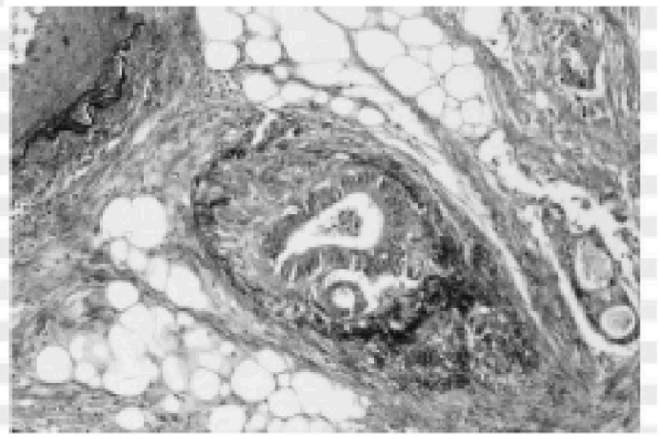


Fig. 4. Grayscale image.

cases. This imbalance can lead to biased models that perform well on the majority class but poorly on the minority class (cancer cases), affecting the predictive accuracy.

Feature Selection and Dimensionality: Identifying the most relevant features from a large pool of variables can be complex. Some features might not contribute significantly to the prediction, while others might be correlated, leading to redundancy and overfitting.

Overfitting and Generalization: Machine learning models might overfit the training data, capturing noise rather than the underlying patterns. This can hinder their ability to generalize and make accurate predictions on new, unseen data.

Interpretability of Models: Complex machine learning models, like deep neural networks, often lack interpretability. In healthcare, interpretability is crucial for understanding the rationale behind predictions, which is necessary for clinical acceptance and decision-making.

Ethical and Legal Concerns: Healthcare data usage raises ethical concerns regarding patient privacy, consent, and the potential for biases within datasets. Compliance with regulations like HIPAA (Health Insurance Portability and Accountability Act) is critical but can also limit data accessibility.

Clinical Implementation: Deploying machine learning models into clinical practice requires validation, integration with existing healthcare systems, and gaining acceptance from healthcare professionals. Ensuring the model's efficacy, reliability, and seamless integration into clinical workflows is a significant challenge.

Addressing these challenges involves collaborations between healthcare professionals, data scientists, and policymakers. It requires advancements in data collection methodologies, model interpretability, and ethical considerations to develop robust and trustworthy predictive models for colorectal cancer.

3. Proposed ranking based colorectal cancer detection model

The proposed ranking based colorectal cancer detection model has the following steps:

- Image Pre-processing
- Image Segmentation
- Feature Extraction using CNN
- Classification based on Ranking Algorithm

3.1. Image pre-processing

In general medical images tend to comprise of noises or loss in quality due to blur which affects the process of disease diagnosis. Noise removal, pre-processing and image enhancement are necessary steps to be applied on the input image prior to the process of machine learning or deep learning models. A detailed survey about the techniques used to clean the image and aids in locating the areas of interest within the given X-ray image was discussed by Kirill Smelyakov et al. [27]. The images that comprise the pixels completely in the shades of gray are termed as Grayscale images. The contrast is represented by mapping the lowest intensity to the black pixels and the highest intensity to the white pixels. Adaptive Gaussian filtering is used to remove the noise during the pre-processing step. Grayscale images have many shades of gray in between the section as shown in Fig. 3 and 4.

3.1.1. Adaptive Gaussian filtering

One of the optimal filters utilized in image processing models is the Gaussian filter. The noise contained in the image can be smoothened out by applying Gaussian filter with a minimal distortion brought in the image. An enhancement to the Gaussian filter known as Adaptive Gaussian filter is also used to de-noise the images [28]. The single dimensional Gaussian filter is represented in Eqn (1).

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \tag{1}$$

where, x denotes the given point and σ is the variance.

The 2-dimensional representation of the same can be represented as in Eqn (2).

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

where, x denotes horizontal axis with distance from origin, and here, y denotes inside vertical axis distance from origin and σ denotes standard deviation of the distribution. The 2-dimensional Gaussian filter can be measured as a stacked version of two 1-dimensional Gaussian filters. The noise removal can be represented as

$$S(x) = F(x) + N(x) \tag{3}$$

where, S(x) denotes the observed signal, F(x) denotes the original signal and N(x) is the independent identical distribution Gaussian noise with zero mean and the variance σ_n^2.

3.1.2. Otsu thresholding

Thresholding is used to segregate the foreground objects from the background objects in the image. Different thresholding algorithms are used to fix the best threshold value. In Otsu's thresholding method shown in Fig. 5, all the likely threshold values are repeatedly updated and lists a measure for the pixels that lie on either sides of the selected threshold. This model ensures that the pixels in the background along with the pixels in the foreground build a decision about the minimum

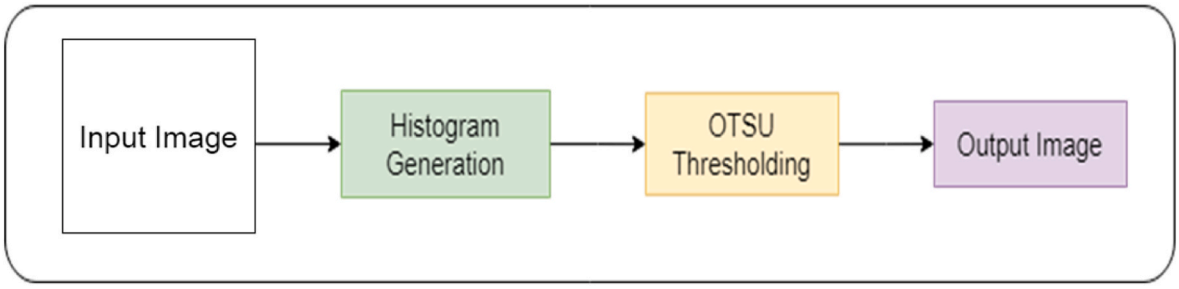


Fig. 5. OTSU thresholding model.

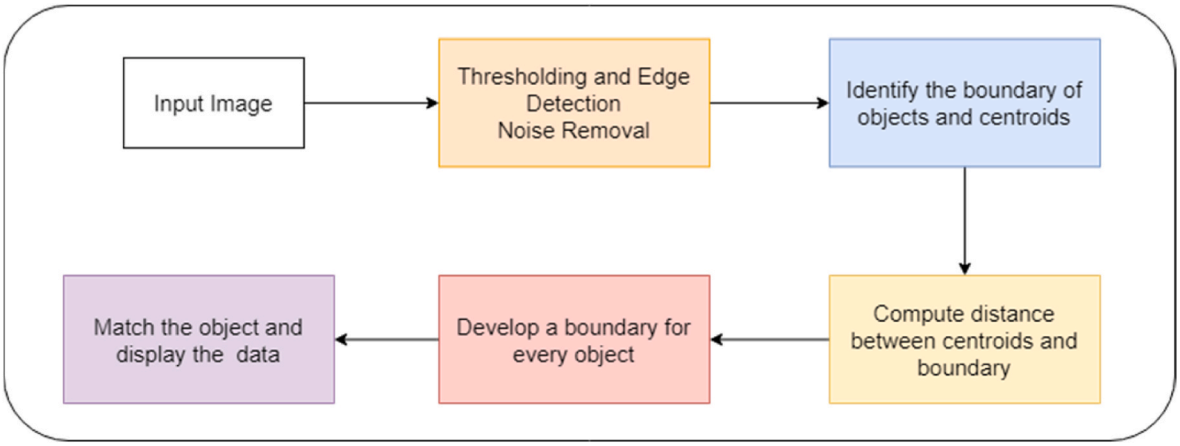


Fig. 6. Watershed segmentation algorithm.

threshold. The optimal value is computed by thinning the total of the weighted group variances. The probability is used to assign weights for the respective groups.

Pi is the probabilities of the observed color values, i = 1, …,K.

$$P_i = \frac{\text{number}\{(r,c)|\text{image}(r,c) = i\}}{(r,c)}$$

r,c denotes the number of row and column count of the image. The variance is calculated as follows:

$$\sigma_w^2 = \omega_b(n) * \sigma_b^2(n) + \omega_f(n) * \sigma_f^2(n)$$

where, $\omega_b(n), \mu_f(n)$ and $\sigma_b^2(n)$ are the weight, average and variance of class Co with intensity 0 to n respectively. $\omega_f(n), \mu_f(n)$ and $\sigma_f^2(n)$ are the weight, average and variance with intensity n+1 to 1 respectively. Otsu thresholding is employed to generate the histogram in the pre-processing phase. This model is developed by imitating the human decision on the disease diagnosis based on their external appearance.

3.2. Image segmentation

Image segmentation is the process of dividing the input image into multiple parts with an aim to minimize the representation of the image as a whole. In general image segmentation is useful to identify unique infection patterns that may support rapid diagnosis, severity assessment, and patient prognosis prediction, but manual segmentations are time-consuming and depend on pathologic expertise. DL-based methods have been explored to reduce the burdens of segmentation; however, their accuracies are limited due to the lack of large, publicly available annotated datasets that are required to establish ground truths. The segmentation of colorectal image is very challenging problem because homogeneity is not present in the region and segmentation of various regions for the diagnosis of cancer is a tedious one. The block diagram of the threshold based salient segmentation method is shown in Fig. 6.

Watershed Image segmentation identifies the objects along with their boundaries and assigns a label to each pixel in an image such that pixels with the similar label share certain characteristics. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Every pixel in a region is similar with respect to some characteristic or property namely color, intensity, or texture. Lines, edges, curves, and other boundary information about the objects are extracted from the MRI image. The affected region is extracted from the colorectal image network using a fully convolutional DL algorithm that employs the preprocessed MRI image as input.

3.3. Feature extraction using CNN

CNN is a class of deep neural networks, most regularly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared-weight architecture of the convolution kernels that shift over input features and provide translation equivariant responses. Counter-intuitively, most CNN are only equivariant, as opposed to invariant, to translation. They have applications in image and videorecognition, recommender systems, image classification, Image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series [29–35].

CNNs are regularizedversions of multilayer perceptronwhere each neuron in one layeris connected to all neurons in the next layer. The "full connectivity" of these networks makes them prone tooverfittingdata. Typical ways of regularization, or preventing overfitting, include: penalizing parameters during training (such as weight decay) or trimming connectivity. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler

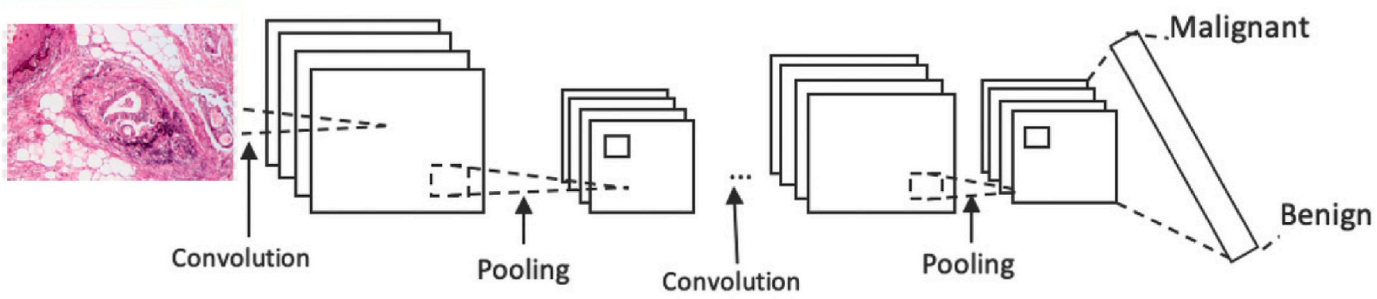


Fig. 7. Proposed CNN Model for colorectal cancer detection.

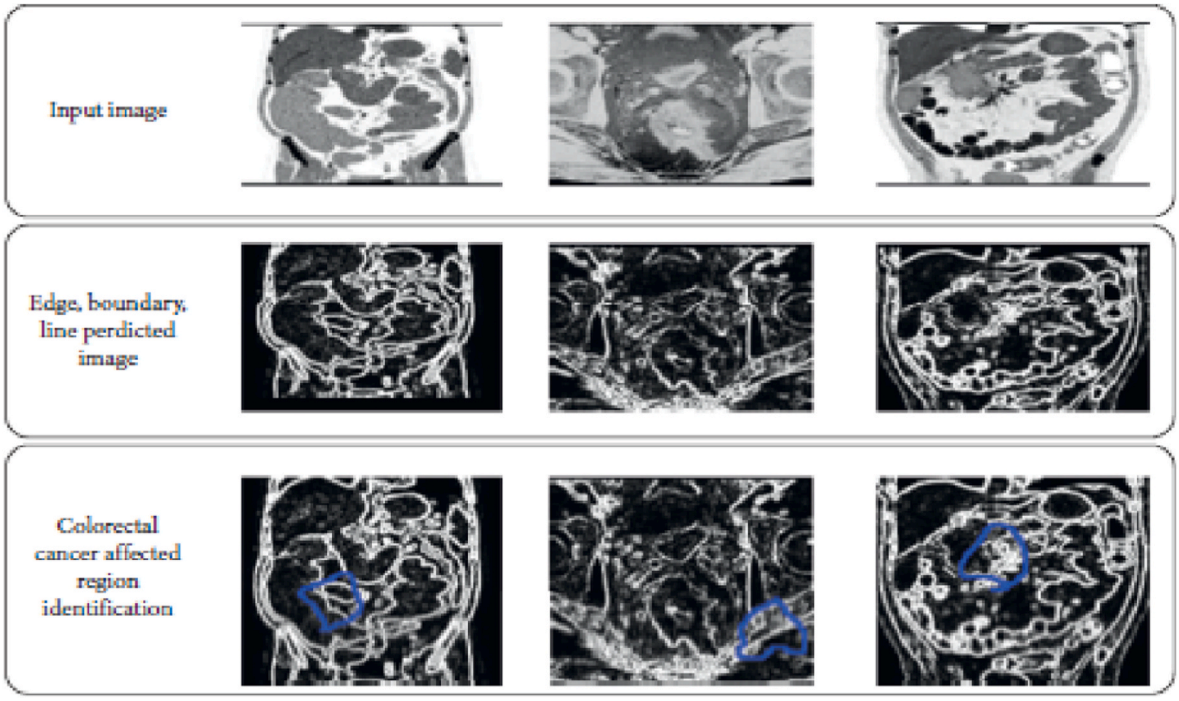


Fig. 8. CNN based region segmented image.

patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.

The proposed colorectal cancer detection models fuses CNN and LSTM simulator modelsto identify the affected tumor tissue is represented in Fig. 7. Tissues are classified into 6stages asStroma, Lympho, Adipose, Complex, Debris and Mucosa [36]. According to a datasheet from cancer.net, the stage of cancer is classified as three types Tumor(T), Node(N), Metastasis(M). Stages are classified based on similarity in multiple tissue types. Combination of these Tumors provides the stage. This is achieved using CNN and LSTM where CNN trains the model to predict the tumor similarity and LSTM helps in identifying. The basic CNN architecture has five layers: the input, convolution, non-linearity (ReLu), pooling, and classification layer (see Fig. 8).

3.3.1. *Input layer*

The initial layer of the network is the input layer. Preprocessing techniques apart from segmenting, normalizing, extracting foreground and back ground objects and converting to gray scale for processing is done to increase the efficiency and reduce the algorithm's computational cost.

3.3.2. *Convolution layer*

The convolution layer carries out two-dimension convolution for three-dimensional input and three-dimensional filter. In case, the size of the input is H × W × C, where height is H, width is W, and the channels is C. Then, the size of the filter is HF × WF × C, where the height and width of the filters is represented by HF and WF respectively. The size of channel for both input and filter is equal. 2D convolution is done along the height and width.

3.3.3. *Non-linearity layer*

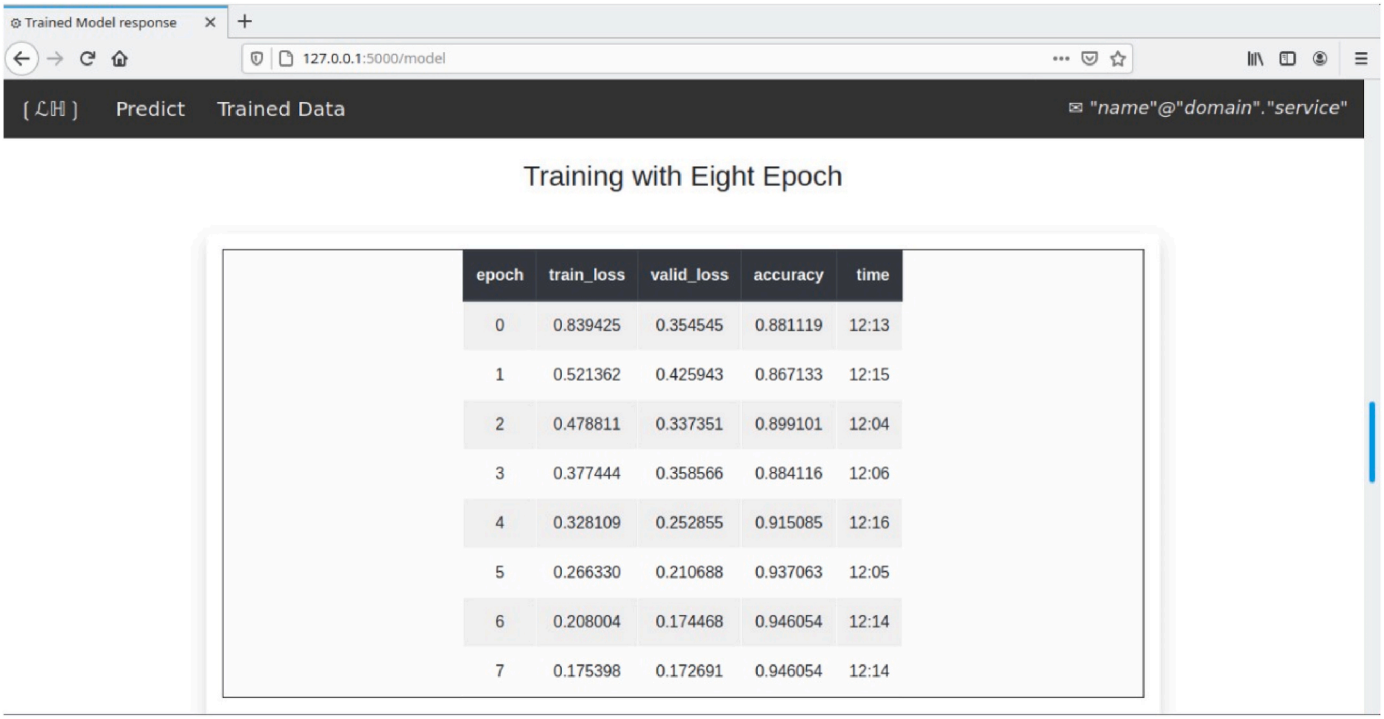
The convolution layer's activation function has to normalize the input image pixels with the help of a non-linear function. Rectified Linear Unit (ReLU) returns zero on receiving a negative input and returns the same value for other inputs.

3.3.4. *Pooling layer*

The pooling layer minimizes the training time and addresses the over-fitting issue by reducing the network dimensionality. The widely used minimization techniques are:

- Setting the arithmetic mean
- Choosing the pixel with maximum value
- Choosing the sum of all the elements

The pooling function substitutes the ReLU layer outputs with a



epoch	train_loss	valid_loss	accuracy	time
0	0.839425	0.354545	0.881119	12:13
1	0.521362	0.425943	0.867133	12:15
2	0.478811	0.337351	0.899101	12:04
3	0.377444	0.358566	0.884116	12:06
4	0.328109	0.252855	0.915085	12:16
5	0.266330	0.210688	0.937063	12:05
6	0.208004	0.174468	0.946054	12:14
7	0.175398	0.172691	0.946054	12:14

Fig. 9. Training data results.

summary of neighboring outputs. It has two advantages:

1. the representation is invariant to small variations in the input
2. reduces the computational load.

3.3.5. *Fully connected (FC) layer*

The FC layer has the total convolution layer and pooling layer. This layer performs high-level analysis to deduce the feature representations from the preceding layers outputs. The number of outputs is the same as the number of classes mentioned in the fully connected layer of a multiclass classification. The softmax function is the most preferred for the task of classification.

3.3.6. *Network structure*

Proposed CNN structure is a 18-layer model. The image input layer inputs the cancer image that is preprocessed. The filter size and the number of filters is set by trial and error. The input is multiplied by a weight matrix by the FC layer. It also adds a bias vector. Softmax layer uses the softmax function, termed as the multiclass generalization of logistic regression.

Consider P(r) is the class prior probability, and P(x|r) is the conditional probability of given class. The probability of sample x belonging to class r is given by equation (4)

$$P(r|x) = \frac{P(x|r)P(x)}{\sum_{k=1}^{R} P(x|k)P(k)} \tag{4}$$

3.4. *Ranking algorithm*

Feature selection is one of the vital steps in colorectal classification. A superior feature is recognized by the following uniqueness [37]:

- A feature which is a part of a single class helps to accurately identify it
- A feature which is a part of every class cannot be used to recognize a particular class
- A feature which is not a part of one or multiple classes can be used for negativity test

Randomized On-Line Matching, a representative of a class of algorithms that exploits a randomized efficient on-line matchingalgorithm and computes maximal matching in bipartite graphs, named the Ranking algorithm, as its basis in a sequential manner. The Ranking algorithm considers that thenodes of one part of the bipartite grapharrive on-line in a sequential way and compute a matching in an on-line fashion. Specifically, the algorithm computes a random per mutationof the nodes in one part of the graph and then considers on-line arrival of the nodes in the other part; each incoming node of the second graph part is matched with the first appropriate node in the permutation of the first graph part. Ranking calculates a maximal matching, as has been proved [38].

Theranking algorithm used in the proposed model is meant to allot a rank for a feature based on its significance towards the target class. If the feature represents the class completely, the score allotted is high. If the feature is a part of all classes, then the probability of considering the feature for ranking in minimal. The following measures are used in the proposed ranking algorithm:

True Positive - TP: suggests that if the example is positive, it has been detected as such
False Negative - FN: suggests that if the example is positive it's detected as negative
True Negative - TN: suggests that if the example is negative it's detected as negative
False Positive - FP: suggests that if the example is negative it's detected as positive

The features are clustered into three groups based on the pearson coefficient, Chi-score and information gain. The pearson coefficient statistical measure to check the relevance of a feature towards the target class [39] and is represented by the following equation:

$$\text{Pearson coefficient} = \frac{COV(x,y)}{\sigma x \sigma y} \tag{5}$$

The Chi-score (CHI_s) is used to compute how much a term is deviating from its dependent class [40] andis computed as:

$$\begin{aligned} CHI_s = {} & t(TP,(TP+FP)) + Prob(P) + t(FN,(FN+TN)) + Prob(P)) \\ & + t(TP,(TP+FP)) + Prob(N) + t(FN,(FN+TN)) + Prob(N)) \end{aligned} \tag{6}$$

where $Prob(N)$ and $Prob(P)$ represent the negative and positive class probability respectively. The information gain is a supervised feature selection method that is used to rank the feature based on the contribution of it in the input data [41]. The proposed ranking algorithm can be represented using the following steps:

Step 1: Rank the features based on TP Score (TPS) – FP Score (FPS)
Step 2: Eliminate the features with minimum FPS
Step 3: Combine the selected features based on the Feature selection algorithm
Step 4: Assign ranking for individual features based on their scores

Ranking Algorithm of the proposed model

Input: F_s= set of features in the input image
Output: R_s – Dominant and high ranked features
BEGIN
for every feature in Feature Set F_s:
    Compute $TPS = \frac{TP}{(TP+FN)}$
    Compute $FPS = \frac{FP}{(TN+FP)}$
    Form a list of features L_s with top 'K' features with high TPS-FPS value
    for each feature 'f' in L_s:
        If Threshold >FPS(f) then:
            Eliminate 'f' from L_s
        Form a feature set F_{s1} based on Pearson Correlation
        Form a feature set F_{s2} based on Information Gain
        Form a feature set F_{s3} based on CHI_s
        Identify the common features from F_{s1}, F_{s2}, F_{s3} and form R_s
Return R_s
END

Ranking algorithm is used to rank each tissue based on the similarity in the image. The tissue with maximum score will be used for predicting the result.

4. Experimental results and discussion

The proposed model has been implemented in a Windows PC with Intel chip set hardware and python programming language. The considered data set is a4-year follow-up data from 334 patients treated for colorectal cancer. From the data set, 284 images were used for training and the test set was of size 50. The data set has been augmented to 5 times by applying simple image preprocessing mechanisms such as image enhancement, image quality adjustment etc. The training results of the proposed model are represented in Fig. 9.

Performance of ML or DL algorithms is usually measured on factors like accuracy, precision and recall as shown in the following formula:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

Table 1
Performance comparison of proposed model.

Classifiers	Precision (%)	Recall (%)	Accuracy (%)
ANN	87	89	88
BPNN	91	91	92
Proposed Ranking Algorithm with CNN	92	93	91

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN}$$

To evaluate the performance of the proposed system, the results are compared with the existing system such as An Ensemble Neural Network (ANN) based colorectal cancer detection proposed by Paladinier al. [42] and Back Propagation Neural Network (BPNN) based colorectal cancer detection suggested by Daniel D. P. et al. [43]. The comparative results are shown in Table 1.

The performance comparison of the existing models and the proposed model for colorectal cancer detection is pictorially represented in Fig. 10.

It is evident from the results that the proposed model outperforms the existing models in the process of cancer diagnosis.

5. Conclusion and future work

Almost half a million people die every year due to colon cancer. Screening for this cancer is effective for prevention as well as early detection. The prediction accuracy in proposed system is better than the existing models. The necessary feature of a good model is to understand the input dataset and the vital features associated with it. It is also a tedious task to manually pick the feature from the dataset that enhances the efficiency of the system. The notable significance of the proposed model is the integration of the CNN and LSTM to achieve a better performance than the existing ones in a faster way. The future scope can be to extend the proposed model to identify multiple types of cancer thereby developing a framework medical image diagnosis.

Declaration of competing interest

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

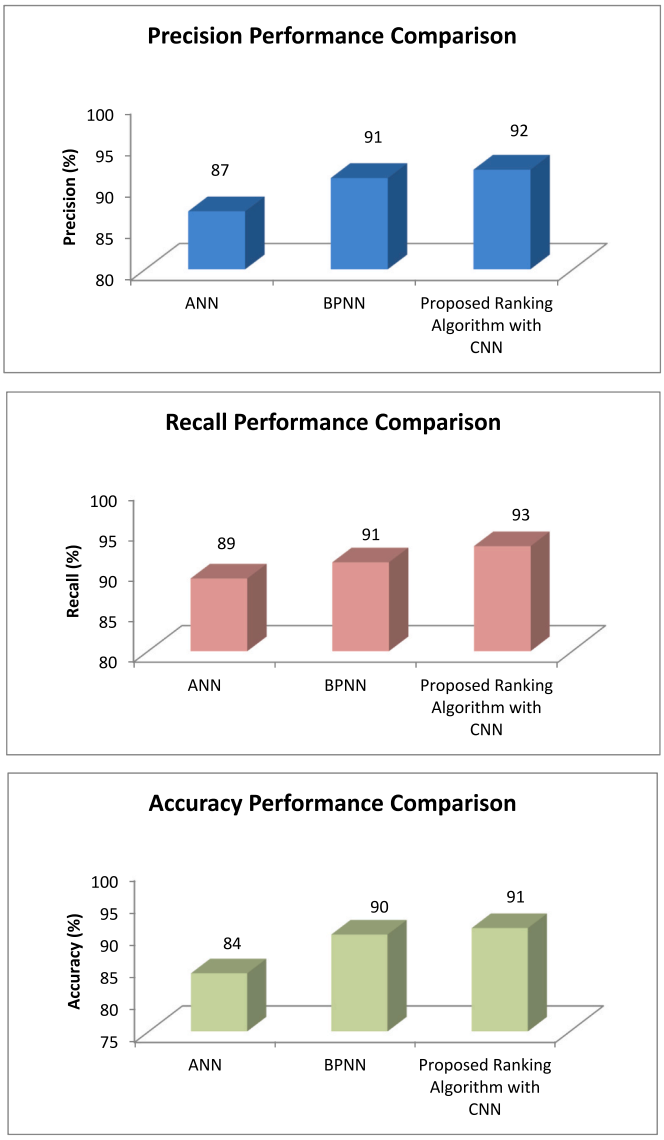


Fig. 10. Performance comparison on a) Precision b) Recall c) Accuracy.

Data availability

No data was used for the research described in the article.

References

[1] P. Dogra, J.D. Butner, Yl Chuang, et al., Mathematical modeling in cancer nanomedicine: a review, Biomed. Microdevices 21 (2019) 40, https://doi.org/10.1007/s10544-019-0380-2.

[2] J. Pascal, E.L. Bearer, Z. Wang, E.J. Koay, S.A. Curley, V. Cristini, Mechanistic patient-specific predictive correlation of tumor drug response with microenvironment and perfusion measurements, Proc. Natl. Acad. Sci. USA 110 (2013) 14266–14271.

[3] KorsukSirinukunwattana, et al., "Locality sensitive deep learning for detection and classification ofNuclei in routine colon cancer HistologyImages", IEEE Trans. Med. Imag. 35 (5) (2016).

[4] W.C.R. Fund, Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective. Colorectal Cancer - from Prevention to Patient Care, AICR, Washington DC, 2007, 2007.

[5] T.K. Noah, B. Donahue, N.F. Shroyer, Intestinal development and differentiation, Exp. Cell Res. 317 (19) (2011) 2702-2710, https://doi.org/10.1016/j.yexcr.2011.09.006.

[6] S. Gout, J. Huot, Role of cancer microenvironment in metastasis: focus on colon cancer, Cancer Microenviron 1 (2008) 69–83.

[7] M. Toğaçar, Disease type detection in lung and colon cancer images using the complement approach of inefficient sets, Comput. Biol. Med. 137 (2021), 104827.

[8] N. Kumar, M. Sharma, V.P. Singh, C. Madan, S. Mehandia, An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images, Biomed. Signal Process Control 75 (2022), 103596.

[9] M. Yildirim, A. Cinar, Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN model: MA_ColonNET, Int. J. Imag. Syst. Technol. 32 (2022) 155–162.

[10] M. Masud, N. Sikder, A.A. Nahid, A.K. Bairagi, M.A. AlZain, A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework, Sensors 21 (2021) 748.

[11] B. I. Strehler and A.S. Mildvan, General theory of mortality and aging, A stochastic model relates observation on aging, physiologic decline, mortality and radiation, Sci. See Saiensu, 132 31-DEC-60.

[12] AR Guru Gokul, et al., Ensembling framework for pneumonia detection in chest X-ray images, in: 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), IEEE, 2022.

[13] Jeroen B. Smaers, Carrie S. Mongle, Anne Kandler, A multiple variance Brownian motion framework for estimating variable rates and inferring ancestral states, Biol. J. Linn. Soc. 118 (2016) 78–94, https://doi.org/10.1111/bij.12765.

[14] Bunil Kumar Balabantaray, Kangkana Bora, Kunio Kasugai, Pallabi Sharma, "Two Stage Classification with CNN for Colorectal Cancer Detection"Oncologie, 2020.

[15] Jiri Prinosil, Malay Kishore Dutta, NamitaSengar, Neeraj Mishra, RadimBurget, "Grading of colorectal cancer using histology images", in: 39th International Conference on Telecommunications and Signal Processing, 2016, 29 June2016.

[16] Nima Tajbakhsh, Suryakanth R. Gurudu, Jianming Liang, Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE, 2015.

[17] Neslihan Bayramoglu, JanneHeikkilä, Transfer learning for cell nuclei classification in histopathology images. European Conference on Computer Vision, Springer, Cham, 2016.

[18] Hawraa Haj-Hassan, Chaddad Ahmad, Youssef Harkouss, ChristianDesrosiers, Matthew Toews, Camel Tanougast, Classificationsof multispectral colorectal cancer tissues using convolution neuralnetwork, J. Pathol. Inf. 8 (2017).

[19] Philipp Kainz, Michael Pfeiffer, Martin Urschler, Segmentationand classification of colon glands with deep convolutional neuralnetworks and total variation regularization, PeerJ 5 (2017), e3874.

[20] Jakub M. Tomczak, Ilse Maximilian, Welling Max, "DeepLearning with Permutation-Invariant Operator for Multi-instanceHistopathology Classification.", 2017 arXiv preprintarXiv:1712.00310.

[21] Hiroshi Yoshida, Yoshiko Yamashita, Eric Cosatto TaichiShimazu, TomoharuKiyuna, Hirokazu Taniguchi, Shigeki Sekine, AtsushiOchiai, Automated histological classification of whole slide imagesof colorectal biopsy specimens, Oncotarget 8 (53) (2017) 9019.

[22] Dmitrii Bychkov, Nina Linder, StigNordling RikuTurkki, PanuE. Kovanen, Verrill Clare, Margarita Walliander, Mikael Lundin, CajHaglund, Johan Lundin, Deep learning based tissue analysispredicts outcome in colorectal cancer, Sci. Rep. 8 (1) (2018) 3395.

[23] Hao Chen, Xiaojuan Qi, Lequan Yu, Dou Qi, Jing Qin, Pheng Ann Heng, DCAN: deep contour-aware networks for object instance segmentation from histology images, Med. Image Anal. 36 (2017) 135–146.

[24] Jianpeng Zhang, Yong Xia, Michael Fulham YutongXie, DavidDaganFeng, Classification of medical images in the biomedicalliterature by jointly using deep and handcrafted visual features, IEEE J. Biomed. Health Inform. 22 (5) (2018) 1521–1530.

[25] ManjuDabass, RekhaVig, and ShardaVashisth. "Five-Grade CancerClassification of Colon Histology Images via Deep Learning." ICCCS2018, Taylor and Francis 2nd International Conference onCommuincation and Computing System.

[26] M. Jayachandran, T. Manivannan, February, "Colorectal cancer detection in MRI images using image processing techniques", Int. J. Eng. Sci. Res. Technol. (2018).

[27] Kirill Smelyakov, et al., Lung X-ray images preprocessing algorithms for COVID-19 diagnosing intelligent systems, COLINS 2022, in: 6th International Conference on Computational Linguistics and Intelligent Systems, vol. I, 2022. Main.

[28] J.C. Uyeda, L.J. Harmon, A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data, Syst. Biol. 63 (2014) 902–918.

[29] J.M. Beaulieu, D.C. Jhwueng, C. Boettiger, B.C. O'Meara, Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution, Evolution 66 (2012) 2369–2383.

[30] Chengwei Xiao, Jiaqi Ye, RuiMáximoEsteves and ChunmingRong, Using Spearman's correlation coefficients for exploratory data analysis on big dataset, CONCURRENCY AND COMPUTATION: practice and experience, Concurrency Pract Ex (2015), https://doi.org/10.1002/cpe.3745. Published online in Wiley Online Library (wileyonlinelibrary.com).

[31] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Stat. 35 (6) (2007) 2769–2794.

[32] J. Pascal, C.E. Ashley, Z. Wang, T.A. Brocato, J.D. Butner, E.C. Carnes, E.J. Koay, C. J. Brinker, V. Cristini, Mechanisticmodeling identifies drug-uptake history as predictor of tumor drug resistance and nano-carrier-mediated response, ACS Nano 7 (2013) 11174–11182.

[33] K.O. Hicks, S.J. Ohms, P.L. van Zijl, W.A. Denny, P.J. Hunter, W.R. Wilson, An experimental and mathematical model for the extravascular transport of a DNA intercalator in tumours, Br. J. Cancer 76 (1997) 894–903.

[34] J. Ciccolini, D. Barbolosi, C. Meille, A. Lombard, C. Serdjebi, S. Giacometti, L. Padovani, E. Pasquier, N. Andre, Pharmacokinetics and pharmacodynamics-

based mathematical modeling identifies an optimal protocol for metronomic chemotherapy, Cancer Res. 77 (2017) 4723–4733.
[35] Z. Wang, R. Kerketta, Y.-L. Chuang, P. Dogra, J.D. Butner, T.A. Brocato, A. Day, R. Xu, H. Shen, E. Simbawa, Theory and experimental validation of a spatio-temporal model of chemotherapy transport to enhance tumor cell kill, PLoS Comput. Biol. 12 (2016), e1004969.
[36] American Cancer Society Key Statistics for Colorectal Cancer, 2017. https://www.cancer.org/cancer/colon-rectal-cancer/about/keystatistics. html, 2018-11-23.
[37] V. Durga Prasad Jasti, GuttikondaKranthi Kumar, M. Sandeep Kumar, V. Maheshwari, PrabhuJayagopal, Bhaskar Pant, AlagarKarthick, M. Muhibbullah, Relevant-based feature ranking (RBFR) method for text classification based on machine learning algorithm, J. Nanomater. 2022 (2022), 9238968, https://doi.org/10.1155/2022/9238968, 12 pages.
[38] Lindsey A. Torre, Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, JoannieLortetTieulent, AhmedinJemal, "Global cancer statistics, 2012.", CA A Cancer J. Clin. 65 (2) (2015) 87–108.
[39] N. Peker, C. Kubat, Application of chi-square discretization algorithms to ensemble classification methods, Expert Syst. Appl. 185 (2021), 115540.
[40] B. Kalaiselvi, M. Thangamani, An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques, Measurement 162 (2020), 107885.
[41] F. Shen, X. Zhang, R. Wang, D. Lan, W. Zhou, Sequential optimization three-way decision model with information gain for credit default risk evaluation, Int. J. Forecast. 38 (3) (2022) 1116–1128.
[42] Emanuela Paladini, et al., Two ensemble-CNN approaches for colorectal cancer tissue type classification, J. Imag. 7 (3) (2021) 51.
[43] Ayush Sharma, SudhanshuKulshrestha, Sibi B. Daniel, Machine learning approaches for cancer detection, Int. J. Eng. Manufact. 8 (2) (2018) 45.