Deep Learning Based Colorectal Cancer (CRC) Tumors Prediction

Rahul Deb Mohalder[1]
Computer Science and Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
cserahul.ku@gmail.com

Kamrul Hasan Talukder[2]
Computer Science and Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
k.h.t@alumni.nus.edu.sg

2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579847

***Abstract*—Cancer detection and prediction using computer assisted systems has become the most leading research area in recent times. It has a big demand in the medical sector for identifying not only cancer but also any diseases detected and predicted from pathological data or images. Colorectal Cancer or Colon Cancer (CRC) detection is also one of them. Because CRC has become a global health issue day by day. In this paper we used a dataset of 10,000 histopathological images with the same dimension of colonic tissue. We used ensemble methods and classifiers for classifying images. We obtained the best accuracy 99% from XGBoost classifier and from others were 98%, 97%, 96%, 92%, 92% and 89% which exactly classifying 523 colon adenocarcinoma images and 477 benign colonic tissue images from 1,000 histopathological images.**

***Index Terms*—Colorectal Cancer, Histopathological Image, Classification, Deep Learning, Ensemble Machine Learning Algorithm**

I. Introduction

Colorectal cancer or colon cancer (CRC) has become a major health issue in the world. From the WHO report CRC is the second major deadliest disease. In 2020 approximately 10 million deaths of cancer and 1 in 6 deaths caused by cancer. This death rate is higher than high income countries to low and medium income countries. CRC is the second most common of deadliest cancer 1.93 million cases are CRC [1]. According to American Cancer Society, at regional periods 56% of CRC patients are diagnosed [2], [3].

CRC incidence has increased both in developing and developed countries day by day and it is a preventable type cancer, during tumor progression time tumor architecture changes and is related to patient prognosis [4]. CRC mainly relies on the position and size of cancer. It may be malignant tissue, benign tissue or non-cancerous tissue. The expected outcome of our research is to perfectly classify the tumors from histopathological images and predict CRC from affected tumors images. There is big demand in the medical sector for a computer assisted system for identifying not only cancer but also any diseases detected and predicted from pathological data or images, which will remove any problem of existing or manual systems. This will be a more convenient and efficient system for identifying CRC in early stages.

Different important application areas of Colorectal Cancer detection such as computer aided diagnosis system [2], medical image classification [5], [6] and intelligent decision framework for cancer diagnosis [7]. To classify, localize and segment affected tumor or cancer affected tumor areas from histology images machine learning techniques are suggested from recent work or research. And for extracting and learning features of subjects deep neural networks (DNNs) are used extensively [4]. Using these techniques relatively small numbers of researchers have tested CRC. To detect cancer at an early stage Kather et al. used decision trees ensemble methods to analyze histology images and they investigated eight types of different colorectal cancer tissue [8].

II. Literature Review

Tumor identification problem from histopathological images researchers use various types of feature extraction methods and prediction models. But they cannot overcome all limitations completely. Some methods and models to solve this problem are described below

A. Histopathological Images Analysis using Deep Learning Techniques for CRC Detection and Prediction

Deep learning (DL) is a part of machine learning. Basically deep learning technique consists of artificial neural networks (ANN). For Image analysis tasks deep learning based techniques are broadly used. Deep learning based techniques accuracy is at an expert level for extracting features from medical images [9]. Medical image analysis using machine learning and deep learning techniques has become more popular day by day. Its accuracy is also at a satisfactory level. Figure 1 shows the DNN which consists of three layers. Input layer, hidden layer and output or output layer. All mathematical operations perform in a hidden layer. The hidden layer mainly stands on the exponential function. In the DNN there are some input values for the input layer and also weight value of each connection neuron. Each neuron performs the activation function (i.e. exponential function) and output data comes from neurons by data standardization. For reducing the cost function value neurons weights are adjusted by Gradient Descent after each iteration.

Bychkov et al. proposed a DL based architecture which can predict CRC from tumor tissues histopathological images [6]. Tumor tissue patterns are so complex. Using classification

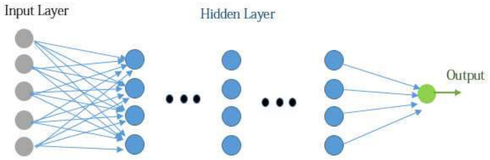

Fig. 1. Deep Neural Network (DNN)).

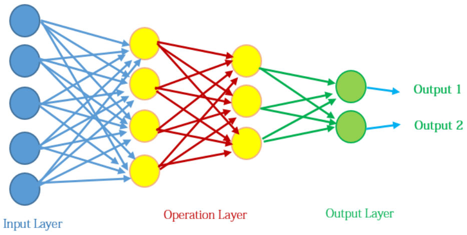

Fig. 2. Neural Network Architecture.

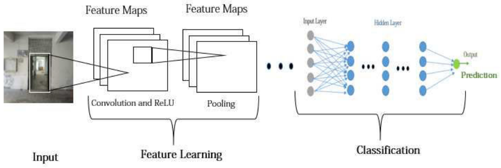

Fig. 3. CNN architecture to classify images.

algorithm features are extracted from tissue images. But they used deep learning methods for feature extraction because feature extraction using classification algorithms takes a long time to extract features. Microscopic digital RGB images split into tiles which contain tumor tissue microarray spots and extract features from those tiles. They trained their model using pre pre-trained dataset and predicted the probability of five year survival. Sirinukunwattana et al. proposed Spatially Constrained CNN method to detect nucleus. For nucleus classification they also proposed a Neighboring Ensemble Predictor. In their proposed system there is no need for any image segmentation process [10]. Colorectal polyps are considered as the primary step of CRC. Korbar et al. proposed a DL method based approaches which can detect five types of polyps [11]. Their proposed image analysis system reduces CRC analysis and diagnosis time.

B. Ensemble Classifiers for CRC Prediction

At the early stage detection and with suitable treatment not only CRC but also any type of cancer can be cured. Stoean et al. proposed system aim was to develop an expert computerized decision system for hospital and diagnosis center which can reduce CRC detection time and predict the level of cancer [7]. For classifying histology images they used ensemble classifiers and those are preamble methods, logistic regression, SVMs, NNs, decision trees and ensemble constructions. Using Wilcoxon rank sum nonparametric test they compared their prediction accuracy. But the logistic regression method result was not at a satisfactory level. Its accuracy was so poor. There was also some misclassified data in all approaches. Dorani et al. proposed genome analysis based CRC risk prediction model. They applied two types of ensemble learning algorithms and those are gradient boosting machine algorithm and random forests algorithm [12]. In the classification process program learns or trains from input data and after the learning process it classifies test data [13]. SVM model mainly used for two group classification problems. It classifies data into two groups which are linearly separable [7], [14]. Neural networks nodes are compared with human brain neurons. But its not fully the human brain identical copy [15]. Each neural network has three parts (figure 2). These parts are called layers.

C. CRC Affected Tumor Detection using Convolutional Neural Network (CNN)

In 2018 Yoon et al. proposed a CNNs based model which can detect affected or not affected tumors from pathological images [4]. Basically CNNs (i.e. Deep Neural Networks) are expert systems for image data analysis. They developed five modifications for detecting tumors from Visual Geometry Group and for measuring performance of CNN models they use modified VGG configuration. They used 400 pathological images which contained both normal and tumor images. And in this task their proposed CNN models performance was best. But their model is suitable for small dataset. For large dataset their model needs some modification and manual interruption for processing. CNN architecture to classify an input image is shown in figure 3.

Feature learning steps are convolution, ReLU and pooling operation. Non Linearity (ReLU) stands for nonlinear operation. If the input image size is too large then the pooling operation reduces parameters. There are three pooling methods (max, sum and average). After the feature learning process the last step is classification and all layers in this step are connected [9], [16]. Classification steps work like the DNN model.

D. CRC Prediction using Intestinal Cancer Prediction Algorithm

Wan et al. proposed a DL based cancer prediction algorithm which can predict CRC in the early stage according to the patient's report and conditions. Their main target was early detection and reducing doctors workloads [17]. Conventional deep learning based systems mainly perform the image processing task and it is so time consuming. But their proposed algorithm analyzes patient reports and shows better results by reducing the time complexity. In their cancer prediction algorithm at first they standardized patient data. The standardization function is

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

Where:

- X is the sample data
- μ is the mean of attributes data
- σ is the standard deviation

After standardizing data they removed redundant attributes which helped to improve the prediction algorithm accuracy. Then they applied a nonnegative matrix factorization method (NMF) for reducing sample data dimension. By reducing sample data dimension they divided data into train data and test data groups. Where a large number of data sets was in the train data group. Finally for the CRC prediction using leveled data they used two models. Deep Belief Network (DBN) based model and Support Vector Machine (SVM) based model. They considered DBN model output as SVM based model input.

E. *Multiclass Texture analysis for CRC Identification*

Tumour histological image contains different types of tissues. But a small number of researchers have worked with multi class problems and there is no work on multi class texture analysis. In 2016, Kather et al. proposed a multiclass tumour texture analysis system for CRC identification [8]. For identifying CRC from histological images they used a large dataset and analyzed eight types of tissue. They analyzed those tissue types for discriminating against different types of CRC tissues. To differentiate into different types of tissue from histological images they used lower order analysis, local texture analysis and local binary pattern process. Kather et al. were the first researchers who used a multiclass texture analysis process for CRC identification. For this reason it is so tough to compare their performance with others' work. But using large dataset their proposed classifier performance was comparatively better than others. More than 25% stage II CRC patient have chance to affect again [18]. A large number of researchers put forward differences between tumour epithelium and stroma tissue by using several types of features matrix [19]. Tumura et al., proposed five visual features that are related to the human perception and these features are used to separate epithelium in the array of CRC such as contrast, roughness, directionality, coarseness and line-likeliness [20]. If it is possible to analysis CRC tissues perfectly then the stage II patient can be survived by additional therapy.

III. Dataset

Dataset was collected from Borkowski et al. which contains 25000 histopathological image dataset of Lung cancer and Colon cancer [21]. We had taken only colon cancers data. In the colon cancer image dataset colon adenocarcinoma images was 5,000 and benign colonic tissue images was 5,000. The dimension (height and width) of each image dataset was 768px (Figure 4). Our selected two types of images were:

a) Adenocarcinoma b) Benign colonic tissue

IV. Proposed Methods

For our prediction work, first we collected histopathological image dataset. In the data acquisition step we collected data from authentic and real data from Borkowski et al [21]. We had used 10,000 colon cancer images from that dataset and each images was in same dimension.

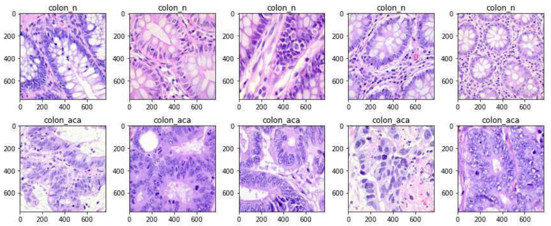

Fig. 4. Datasets

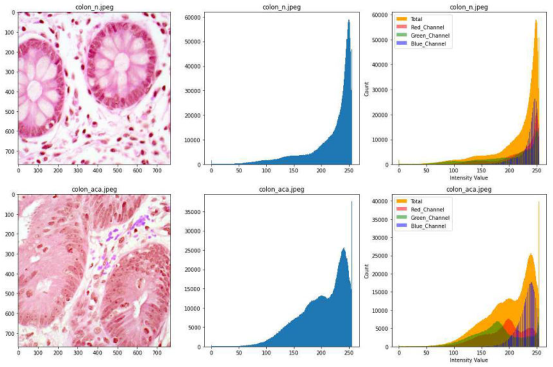

Fig. 5. Input Images Intensity Analysis.

A. *Exploratory Data Analysis*

In the exploratory data analysis process we analyzed our data in three ways. They were image intensity distribution per class, image analysis through comet-tail graph and clustering analysis. By using these techniques we classified histopathological images based on tumor or not.

B. *Image Pre-Processing*

After completing the exploratory data analysis process we filtered images using Local Binary Pattern (LBP), Hog Filter and Gabor Filter techniques. After filtering images we identified and differentiated affected and normal tumors from tumor images.By K-means and Region Based Convolutional Neural Network (R-CNN) method we segmented images. These methods were used for labeling or categorizing unlabeled data. After segmenting objects from images we checked objects and analyzed each object. Figure 5 shows the histogram and image intesity graph of colon adenocarcinoma and colinic bengin tissue images.

C. *Train-Test Split*

We splitted our data into train and test data. Model learned from those training data which make the model generalized for other data because it contains the known output. Using a test data set we tested the model for prediction.

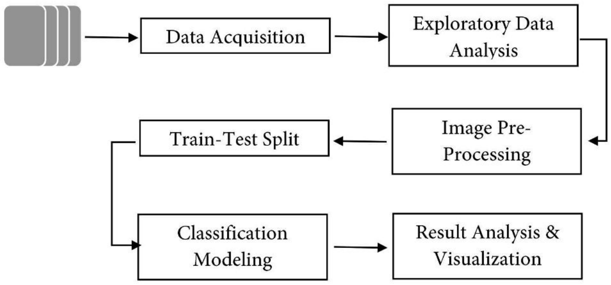

Fig. 6. System Architecture.

D. Classification Modeling

We classified our model using Support Vector Machine (SVM), Decision Tree, Random Forest, Ensemble Method, CNN and Transfer Learning methods. SVM is a supervised machine learning algorithm. For classifying tumor histopathological images we used SVM. Decision Tree is a simple decision making diagram. But for prediction or classification task decision trees are more powerful tools. Applying decision tree technique on tumor images we classified tumors and predict. Using a number of Decision Trees the Random Forest model is created. Using Random Forest model we can classify our data (images) more specifically. To better understand the tumor image we used Ensemble methods, Ensemble methods technique is the combination of several base model of machine learning technique. And this method gave us optimal and best predictive solution. Convolutional Neural Network (CNN) is the best neural network technique for Deep Learning based work. CNN technique was used for analyzing complex data. By CNN technique we analyzed our complex tumor images for identifying abnormal or suspicious tumor patterns. Transfer learning techniques work in the analysis process based on the knowledge gain process. But transfer learning techniques are appropriate when we have insufficient data.

V. System Architecture

Figure 6 illustrates our proposed system architecture. In the first step colon histopathological images are checked and classified. By the exploratory data analysis process images are also analysed by image intensity analysis, analysis through comet-tail graph and clustering analysis. In the image pre-processing step images are processed by LBP, Hog Filter and segmented by K-means, R-CNN. We set a ratio to divide the whole data into a training and testing set. Then the classification modeling step. In this step we classified our model by SVM, Decision Tree, Random Forest, CNN and Transfer learning methods. We trained our model using the training dataset and verified the model by the testing dataset.

VI. Results and Visualization

In this research a dataset of 10,000 colon cancer histopathological images was used. In our dataset of 10,000 images there were 5,000 colon adenocarcinoma and 5,000 of colonic benign tissues histopathological image. We set 90% data for training our model and 10% data for verifying or testing our model. For classifying the model we used seven classifiers. We got the best classification accuracy from the XGBoost classifier and its accuracy was 99%. And others classifiers accuracy was within 98% to 89% (Table I). Figure 7 illustrates the ML algorithm comparisons output. We also applied 10 random cross validation method in our model. For analyzing result and measuring accuracy or performance (2) of our model we used Precision (3), Recall (4) and F1-score (5) technique.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{5}$$

Where:

- TP is the number of true positives
- TN is the number of true negatives
- FP is the number of false positives
- FN is the number of false negatives

TABLE I
Accurecy and Loss from seven Algorithoms and Classifiers

Algorithoms or Classifiers	Accurecy	Loss
Ensemble Machine Learning algorithm(XGBoost)	99%	0.0046
Random Forest(RF)	98%	0.0046
k-Nearest Neighbors (KNN)	97%	0.0036
Decision Trees (DT)	96%	0.0086
Linear Discriminant Analysis (LDA)	93%	0.0080
Support Vector Machine (SVM)	92%	0.0132
Logistic Regression (LR)	89%	0.0178

The Table II illustrates the performance of our model. XGBoost classifiers performance is derived from the average value of precision, recall, f1-score value.

TABLE II
Persormane Measures of XGBoost Classifier

	Precision	Recall	F1-Score	Support
0	1.00	0.98	0.99	523
1	0.98	1.00	0.99	477
accuracy			0.99	100
macro avg	0.99	0.99	0.99	1000
weighted avg	0.99	0.99	0.99	1000

Figure 8 shows the confusion matrix output having test data for colon adenocarcinom which class level is 0 and colonic

TABLE III
PERFORMANCE COMPARISON BETWEEN VARIOUS COLON CANCER PREDICTION AND DETECTION TECHNIQUES

Reference	Models and Methods	Dataset	Pre-processing	Results
Our Proposed Model	Proposed method with XGBoost classifier	10,000 histopathological image with 768pxX768px dimention	Processed images using traditional methods	99% from XGBost, 98% from RF, 97% from KNN, 96% from DT, 92% from LDA and SVM, 89% from LR
Kather et al [8]	Ensemble Classifiers	5000 CRC histopathological images with eight types of tissue	Separate images in eight tissue groups with 625 images and convert images into 150X150 pixels tiles.	98.6% accuracy in traditional methods and 87.4% accuracy in eight type tissue, which are separated in multiclass
Yoon et al. [4]	CNNs, VGG	29 tumour and 28 normal CRC histology images from South Korea National Cancer Center.	Cropped tumour and normal images into smaller sizes into 256X256 pixels and they got a total 10280 images set after processing.	94.29% accuracy with 0.365 loss value in the first experiment and 93.48% accuracy with 0.4464 loss value in the second experiment.
Stoean et al. [7]	Ensemble Classifiers	368 CRC patients diagnosed data from Craiova University of Medicine and Pharmacy, Romania.	Segment clinical images and extract feature in a specified regulation.	Perfectly classify CRC samples into two categories for length of stay.
Wan et al. [17]	DL, DBN, SVM	Two years clinical CRC patients data from Universities of Chinese Medicine in Nanjing (Feb, 2014-Feb, 2016).	Extract k-dimensional features from m-dimensional features by applying nonnegative factorization method in tumour images.	91% accuracy, 0.5 precision, 0.11 recall, 0.18 F1-Score and 0.88 FNR.

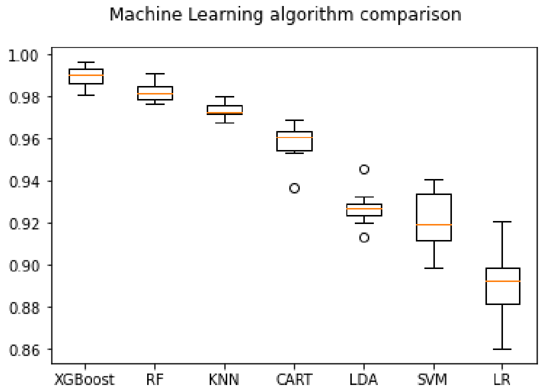

Fig. 7. Machine Learning Algorithm comparison

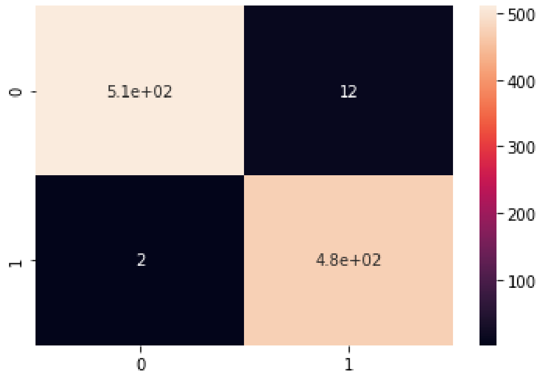

Fig. 8. Confussion Matrics

benign tissue which class level is 1. This confusion matrix also represents the XGBoost classifier output for the test dataset.

VII. DISCUSSION

We compared our results with other researchers' work. Table III illustrates the comparison between different colon cancer prediction or detection techniques.

VIII. CONCLUSION

Colorectal Cancer Tumour prediction from histopathological images is a long standing issue in the medical field. The expected outcome of this area is to perfectly classify the tumours from histopathological image and predict CRC from affected tumour images. There are many convenient and efficient systems for identifying CRC in early stages. But most of them are human and diagnosis center oriented systems. In this paper, we proposed a deep learning based prediction model which is able to classify and predict tumor images from mixed datasets. Our best accuracy was 99% and other methods accuracy for our model was 98% to 89%. Our model's limitation is that it can only classify colon cancer image data and noise free data perfectly. In future we want to work with other deadliest types of cancer. And we also want to tune our model more flexible for any kinds of data (textual or image or nosiy).

References

[1] "Cancer," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] J. Malik, S. Kiranyaz, S. Kunhoth, T. Ince, S. Al-Maadeed, R. Hamila, and M. Gabbouj, "Colorectal cancer diagnosis from histology images: A comparative study," *arXiv preprint arXiv:1903.11210*, 2019.

[3] R. Parveen, S. S. Rahman, S. A. Sultana, and Z. H. Habib, "Cancer types and treatment modalities in patients attending at delta medical college hospital," *Delta Medical College Journal*, vol. 3, no. 2, pp. 57–62, 2015.

[4] H. Yoon, J. Lee, J. E. Oh, H. R. Kim, S. Lee, H. J. Chang, and D. K. Sohn, "Tumor identification in colorectal histology images using a convolutional neural network," *Journal of digital imaging*, vol. 32, no. 1, pp. 131–140, 2019.

[5] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P. E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[6] A. K. Jain and S. Lal, "Feature extraction of normalized colorectal cancer histopathology images," in *Ambient Communications and Computer Systems*. Springer, 2019, pp. 473–486.

[7] R. Stoean, C. Stoean, A. Sandita, D. Ciobanu, and C. Mesina, "Ensemble of classifiers for length of stay prediction in colorectal cancer," in *International Work-Conference on Artificial Neural Networks*. Springer, 2015, pp. 444–457.

[8] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner, "Multi-class texture analysis in colorectal cancer histology," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[9] R. Mohanasundaram, A. S. Malhotra, R. Arun, and P. Periasamy, "Deep learning and semi-supervised and transfer learning algorithms for medical imaging," in *Deep Learning and Parallel Computing Environment for Bioengineering Systems*. Elsevier, 2019, pp. 139–151.

[10] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[11] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep learning for classification of colorectal polyps on whole-slide images," *Journal of pathology informatics*, vol. 8, 2017.

[12] F. Dorani, T. Hu, M. O. Woods, and G. Zhai, "Ensemble learning for detecting gene-gene interactions in colorectal cancer," *PeerJ*, vol. 6, p. e5854, 2018.

[13] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.

[14] D. Fradkin and I. Muchnik, "Support vector machines for classification," *DIMACS series in discrete mathematics and theoretical computer science*, vol. 70, pp. 13–20, 2006.

[15] L. Burke and J. P. Ignizio, "A practical overview of neural networks," *Journal of Intelligent Manufacturing*, vol. 8, no. 3, pp. 157–165, 1997.

[16] W. H. L. Pinaya, S. Vieira, R. Garcia-Dias, and A. Mechelli, "Convolutional neural networks," in *Machine Learning*. Elsevier, 2020, pp. 173–191.

[17] J.-J. Wan, B.-L. Chen, Y.-X. Kong, X.-G. Ma, and Y.-T. Yu, "An early intestinal cancer prediction algorithm based on deep belief network," *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.

[18] A. Huijbers, R. Tollenaar, G. v Pelt, E. Zeestraten, S. Dutton, C. McConkey, E. Domingo, V. Smit, R. Midgley, B. Warren *et al.*, "The proportion of tumor-stroma as a strong prognosticator for stage ii and iii colon cancer patients: validation in the victor trial," *Annals of oncology*, vol. 24, no. 1, pp. 179–185, 2013.

[19] J. K. Shuttleworth, A. G. Todman, R. N. Naguib, B. M. Newman, and M. K. Bennett, "Colour texture analysis using co-occurrence matrices for classification of colon cancer images," in *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No. 02CH37373)*, vol. 2. IEEE, 2002, pp. 1134–1139.

[20] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, man, and cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[21] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (lc25000)," Dec 2019. [Online]. Available: https://arxiv.org/abs/1912.12142