Msi And Mss Classification Of Gastrointestinal Cancer Using CNN

M A Anu Dharshini[1], C Jasphin[2]
[1, 2] Dept of CSE
[1, 2] Arunachala College of Engineering for Women

Abstract- *Gastrointestinal and Colorectal cancers are treated with chemotherapy and its other forms which are not able to provide higher survival rates. Immunotherapy is increasingly becoming popular due to its promising response especially to mutated tumors such as MicroSatellite Instability (MSI) cancers with deficient DNA Mismatch-Repair system (DMMR). Generally, 85% of all the cases related to gastrointestinal and colorectal cancers have proficient DNA Mismatch-Repair system (PMMR) which are also called MicroSatellite Stability (MSS). Only about 15% of the gastrointestinal and colorectal cancer patients have deficient DNA Mismatch-Repair system causing MicroSatellite Instability (MSI) in their tumors. While Immunotherapy responds well to patients with MSI tumors, it is resistant to MSS tumors. Hence, it's important to classify MSI vs. MSS tumors so that appropriate treatment can be given to the patients. Clinically MSI cancers are difficult to be detected after stage III due to their sensitivity to pembrolizumab inhibitors . In this work, deep learning based CNN approach is detailed that can accurately classify MSI vs. MSS cancers using histological images which are derived from Formalin-Fixed Paraffin-Embedded (FFPE).*

Keywords- MicroSatellite instabilityt, Tumors, Gastrointestinal

I. INTRODUCTION

Deep Learning which has emerged as an effective tool for analyzing big data – uses complex algorithms and artificial neural networks to train machines/computers so that they can learn from experience, classify and recognize data/images just like a human brain does. Within Deep Learning, a Convolutional Neural Network or CNN is a type of artificial neural network, which is widely used for image/object recognition and classification. Deep Learning thus recognizes objects in an image by using a CNN. CNNs are playing a major role in diverse tasks/functions like image processing problems, computer vision tasks like localization and segmentation, video analysis, to recognize obstacles in self-driving cars, as well as speech recognition in natural language processing. As CNNs are playing a significant role in these fast-growing and emerging areas, they are very popular in Deep Learning.

Gastrointestinal cancer is cancer that develops along the GI tract (also called the digestive tract). The GI tract starts at the esophagus the tube that takes food from the mouth to the stomach) and ends at the anus where waste exits the body. Primary GI cancer starts growing in the GI tract. Metastatic gastrointestinal cancers originate in the GI tract, but spread to the other parts of the body. There were an estimated 4.8 million new cases of gastrointestinal (GI) cancers and 3.4 million related deaths, worldwide, in 2018. GI cancers account for 26% of the global cancer incidence and 35% of all cancer-related deaths. We investigated the global burden from the 5 major GI cancers as well as geographic and temporal trends in cancer-specific incidence and mortality.

As per IARC(International Agency for Research on Cancer) 1 in 5 people develop cancer. Among all cancer related deaths Gastrointestinal Cancer constitutes to 35% of global cancer related deaths. Computer Vision was used to detect cancer tumors through histological images which drastically cut down both the time and money to carry out conventional testing methods. Gastrointestinal cancers account for 26% of the global cancer incidence burden and 35% of all cancer-related deaths; in 2018, there were an estimated 4.8 million new cases and 3.4 million related deaths worldwide. The researchers analysed patterns and trends in the incidence of and mortality from cancers of the oesophagus, stomach, colorectum, liver, and pancreas.

Microsatellite is defined as the rudimentary repetitive sequence of the Deoxyribonucleic acid (DNA). DNA comprises of many microsatellites. DNA Mismatch Repair (MMR) is a system which monitors the replication process of microsatellites and DNA, if it finds any error in the DNA recombination and replication it performs repair with the help of MMR proteins. Failure of MMR leads to unstable microsatellites/DNA which is the genesis of cancer. Based on global genomic status cancer tumor is classified into 'Microsatellite instable' (MSI) and 'Microsatellite Stable' (MSS) tumor. High amount of instability in tumor classifies it as MSI-H and it can be inherited, in which the immune cells

are shut off from fully doing their job. By using 'Immunotherapy' MSI-H can be cured .In MSS the DNA in tumor cell has the same number of microsatellites that of a healthy cell, this can be cured by 'radiation' and 'chemotherapy'-treatments which are opposite to immunotherapy. 26.4% of gastrointestinal cancer patients are classified as MSI-H and the rest i.e. 73.6% as MSS .Therefore, detection of MSI or MSS of cancer has the same significance as detection of cancer to give appropriate treatment.

MSI stands for "Microsatellite Instable." MSI-H means that there is a high amount of instability in a tumor. MSI-H results when genes that regulate DNA (called Mismatch Repair Genes) don't work correctly. Mismatch Repair Genes (MMR) work like genetic "spell checkers" by correcting errors in DNA as cells divide, similar to how "spell checkers" correct typos on a computer. When MMR genes stop functioning at their highest potential, areas of DNA could start to become unstable due to the errors. An MSI screening test looks for changes in the DNA sequence between normal tissue and tumor tissue and can identify whether or not there is a high amount of instability, which is called MSI-High.

II. LITERATURE SURVEY

The herpesvirus, polyomavirus, papillomavirus, and retrovirus families are associated with breast cancer. More effort is needed to assess the role of these viruses in the detection and diagnosis of breast cancer cases in women. The aim of this paper is to propose an efficient segmentation and classification system in the Mammography Image Analysis Society (MIAS) images of medical images. Segmentation became challenging for medical images because they are not illuminated in the correct way. The role of segmentation is essential in concern with detecting syndromes in human. This research work is on the segmentation of medical images based on Intuitionistic Possibilistic Fuzzy C-Mean (IPFCM) Clustering. Intuitionist Fuzzy C-Mean (IFCM) and Possibilistic Fuzzy C-Mean (PFCM) algorithms are hybridised to deal with problems of fuzzy c-mean. The introduced clustering methodology, in this article, retains the positive points of PFCM which helps to overcome the problem of the coincident clusters, thus the noise and less sensitivity to the outlier. The IPFCM improves the fundamentals of fuzzy c-mean by using intuitionist fuzzy sets.

Deep learning methods, and in particular convolutional neural networks have led to an enormous breakthrough in a wide range of computer vision tasks, primarily by using large-scale annotated datasets. However, obtaining such datasets in the medical domain remains a challenge. In this paper, present methods for generating synthetic medical images using recently presented deep learning Generative Adversarial Networks (GANs). Furthermore, show that generated medical images can be used for synthetic data augmentation, and improve the performance of CNN for medical image classification. Novel method is demonstrated on a limited dataset of Computed Tomography (CT) images of 182 liver lesions 53 cysts, 64 metastases and 65 hemangiomas. First exploit GAN architectures for synthesizing high quality liver lesion ROIs. Present a novel scheme for liver lesion classification using CNN. Finally, train the CNN using classic data augmentation and synthetic data augmentation and compare performance. In addition, explore the quality of synthesized examples using visualization and expert assessment. The classification performance using only classic data augmentation yielded 78.6% sensitivity and 88.4% specificity. By adding the synthetic data augmentation the results increased to 85.7% sensitivity and 92.4% specificity.

Image segmentation is an important task in many medical applications. Methods based on convolutional neural networks attain state-of-the-art accuracy; however, they typically rely on supervised training with large labeled datasets. Labeling medical images requires significant expertise and time, and typical hand-tuned approaches for data augmentation fail to capture the complex variations in such images. Present an automated data augmentation method for synthesizing labeled medical images. Demonstrate the task of segmenting Magnetic Resonance Imaging (MRI) brain scans. Method requires only a single segmented scan, and leverages other unlabeled scans in a semi-supervised approach. Learn a model of transformations from the images, and use the model along with the labeled example to synthesize additional labeled examples. Each transformation is comprised of a spatial deformation field and an intensity change, enabling the synthesis of complex effects such as variations in anatomy and image acquisition procedures. Semantic image segmentation is crucial to many biomedical imaging applications, such as performing population analyses, diagnosing disease, and planning treatments. When enough labeled data is available, supervised deep learning-based segmentation methods produce state of-the-art results.

Microsatellite instability determines whether patients with gastrointestinal cancer respond exceptionally well to immunotherapy. However, in clinical practice, not every patient is tested for MSI, because this requires additional genetic or immunohistochemical tests. Here we show that deep residual learning can predict MSI directly from H&E histology, which is ubiquitously available. This approach has the potential to provide immunotherapy to a much broader subset of patients with gastrointestinal cancer.

Deep learning can mine clinically useful information from histology. In gastrointestinal and liver cancer, such algorithms can predict survival and molecular alterations. Once pathology workflows are widely digitized, these methods could be used as inexpensive biomarkers. However, clinical translation requires training interdisciplinary researchers in both programming and clinical applications.

Due to privacy concerns, of large public databases of medical pathologies is a well-known and major problem, substantially hindering the application of deep learning techniques in this field. In this article, investigate the possibility to supply to the deficiency in the number of data by means of data augmentation techniques, working on the recent Kvasir dataset Pogorelov et al., 2017of endoscopical images of gastrointestinal diseases. The dataset comprises 4,000 colored images labeled and verified by medical endoscopists, covering a few common pathologies at different anatomical landmarks: Z-line, pylorus and cecum.Show how the application of data augmentation techniques allows to achieve sensible improvements of the classification with respect to previous approaches, both in terms of precision and recall. data augmentation can provide a valid palliative to the small dimension of the above mentioned dataset, proving that the problem of automatic diagnosing of gastrointestinal diseases from images can be successfully addressed by means of deep learning algorithm. Data augmentation is a key technique of machine learning. It consists in increasing the number of data, by artificially synthesizing new samples from existing ones, usually via minor perturbations. For instance, in the case of images, typical operations are rotation, lighting modifications, rescaling, cropping and so on; even adding random noise can be seen as a form of data augmentation.

Data augmentation is an effective technique for improving the accuracy of modern image classifiers. However, current data augmentation implementations are manually designed. Describe a simple procedure called Auto Augment to automatically search for improved data augmentation policies. Implementation, have designed a search space where a policy consists of many sub policies, one of which is randomly chosen for each image in each mini-batch. A sub-policy consists of two operations, each operation being an image processing function such as translation, rotation, or shearing, and the probabilities and magnitudes with which the functions are applied. Use a search algorithm to find the best policy such that the neural network yields the highest validation accuracy on a target dataset. Method achieves state-of-the-art accuracy on CIFAR-10, CIFAR-100, SVHN, and ImageNet (without additional data). On Image Net, attain a Top-1 accuracy of 83.5% which is 0.4% better than the previous record of 83.1%. On CIFAR-10, achieve an error rate of 1.5%, which is 0.6% better than the previous state-of-the art. Augmentation policies find are transferable between datasets. The policy learned on ImageNet transfers well to achieve significant improvements on other datasets, such as Oxford Flowers, Caltech-101, Oxford-IIT Pets, FGVC Air craft, and Stanford Cars.

Automatic polyp detection by using a model of polyp appearance in the context of the analysis of colonoscopy videos. Method consists of three stages: region segmentation, region description and region classification. The performance of region segmentation method guarantees that if a polyp is present in the image, it will be exclusively and totally contained in a single region. The output of the algorithm also defines which regions can be considered as non-informative. Define as region descriptor the novel Sector Accumulation-Depth of Valleys Accumulation (SA-DOVA), which provides a necessary but not sufficient condition for the polyp presence. Finally, classify segmented regions according to the maximal values of the SA-DOVA descriptor. Preliminary classification results are promising, especially when classifying those parts of the image that do not contain a polyp inside During the last decades there is a trend that consists of developing intelligent systems for medical applications. Intelligent systems are currently being used to assist in other medical interventions.

III. PROPOSED SYSTEM

Proposed method using CNN algorithm. Convolutional neural networks introduced by are special type of feed-forward artificial neural networks which is inspired from visual cortex. A small region in the brain called Visual cortex is a region of cells that are sensitive to specific regions of visual field. That is, few neurons in the visual cortex fires when exposed to vertical edges while few fire when exposed to horizontal layers. Few fire and exposed to diagonal edges and that is the motivation behind convolutional neural network. If an image of 200 x 200 x 3 pixels is fed to a fully connected neural network, around 120 thousand bits are required at the first hidden layer itself which require a lot of parameters. Basically in a convolutional neural network each neuron in one layer is connected with another layer of the network that contains a small region of the layer before it.

This topology results in a fewer weights between neurons as the number of connections between layers are low. CNN considers small segments of the image where these segments/patches are known as features or filters. By finding a matching feature in roughly the same positions in two images, CNN improves on learning the similarity between the whole image matching schemes. In convolution layer, one by one feature is taken and moved it through the entire image. While

moving filter CNN's multiplies the pixel value of the image with that of the corresponding pixel value of the filter and can be added and dividing by the total number of pixels to get the output. Generally a convolutional neural network has three layers.

A convolution layer, pooling layer and towards the end a fully connected layer. Convolutional neural networks or CNNS can do some pretty interesting things when they are fed with a bunch of pictures. For instance when the face images are given as an input, the convolution neural networks learns some for the features such as edges, dots and spots. These multi-layer neural networks learn these edges or gradients in the initial layers and the second layer learns some of the parts of objects such as eyes, noses and mouths. The third layer learns objects such as faces.

Convolution is a measure of overlap between two functions as one slides over the other. Mathematically it's a sum of products the standard convolution operation is slow to perform however it can speed up with an alternative method called depth wise separable convolution. A scalar is returned from a regular convolution operation that computes the input's and kernel's sum of products. This operation is continued by sliding the kernel over the input. The concern now is with the cost of this convolution operation which has a number of multiplications required.

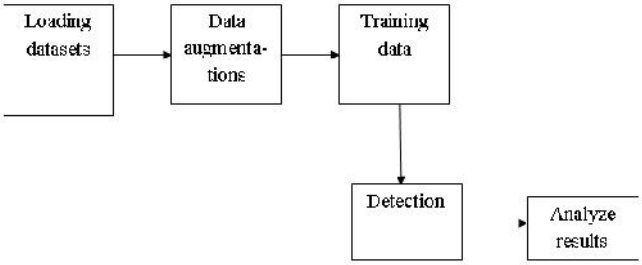


Fig 1 Overview of the Classification of Gastrointestinal Cancer

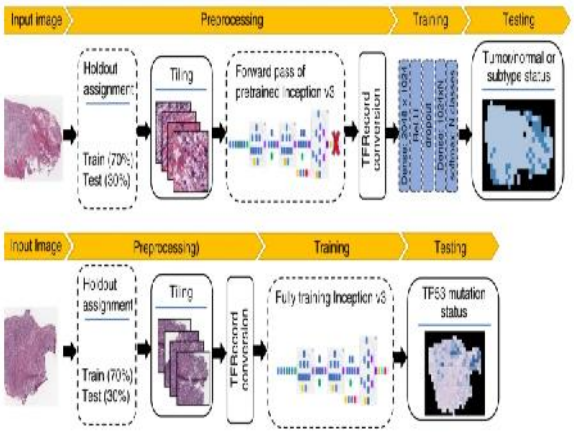


Fig 2 System Architecture of Proposed System

Above diagram explained about system architecture. Developed a CNN architecture to classify slides from TCGA by tumor/normal status, using a neural network that feeds the last fully connected layer of an Inception v3-based CNN pretrained on Image Net into a fully connected layer with 1024 neurons and a related architecture for mutation classification described in sections below. The two final fully connected layers of the tumor/normal CNN were trained on tiles of size 512 × 512 from WSIs.

Due to insufficient FFPE normal WSIs in TCGA, for this task, we only used flash-frozen samples. Trained this model separately on slides from 19 TCGA cancer types having numbers of slides ranging from 205 to 1949. In all, 70% of the slides were randomly assigned to the training set and the rest were assigned to the test set.

The model parameters are configured firstly, data enhancement is selected, training the model with pathological sections of gastrointestinal cancer, and then the test picture is input into the inference model to obtain a masked predicted image and an AP value. Then adjust the parameters according to the results, use different data enhancement methods, retrain the model, make detection, and then choose the best parameter configuration and data augmentation method.

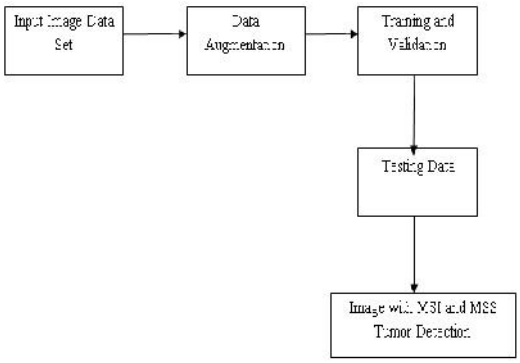


Fig 3 Block Diagram of Proposed System

IV. MODULES DESCRIPTION

Database

The dataset used in this work comes from which has histological images of gastrointestinal cancer that can be used to classify MSI Vs MSS. The original dataset The dataset used in this work comes from which has histological images of gastrointestinal cancer that can be used to classify MSI Vs MSS.

The patients were categorized by specialists into MSI and MSS categories so that dataset has labels.

The dataset is divided into training, validation and testing in the ratio of 80%, 10% and 10% respectively. The total training images 153849.Total validation images: 19230 .Total testing images: 19233

Data augmentation

In order to make the model more robust, the training image data can be augmented. This will address any issue of overfitting and the model gets more generalized. The input training images are rescaled and then rotated by 45 degrees. Then the images are shift by 20% both vertically and horizontally. Later the images are flipped horizontally and zoomed by 50%. This overall process augments the images and improves the validation accuracy.

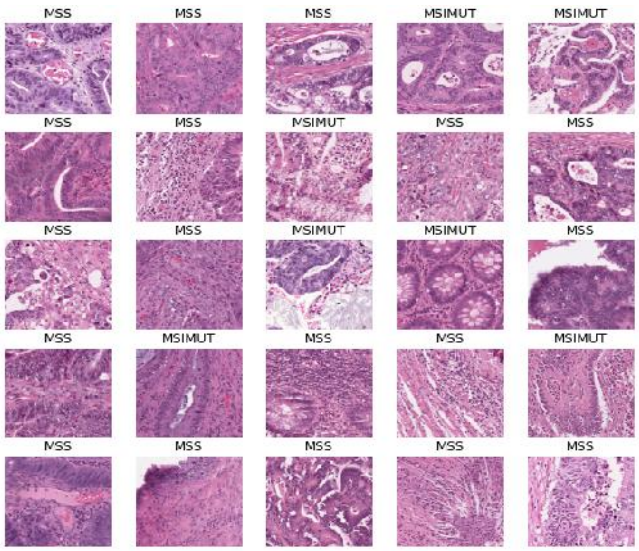
Training and Validation

The training of the pre-trained Xception network is done with histological training images of the dataset. The overall training images are 153,849 and the overall validation images are about 19230. The overall program execution was done in Google Colab which provides free GPU access. The training was done for up to 5 epochs which is sufficient as the initial layers of the Xception network are already pretrained to identify edges, gradients, textures, etc.

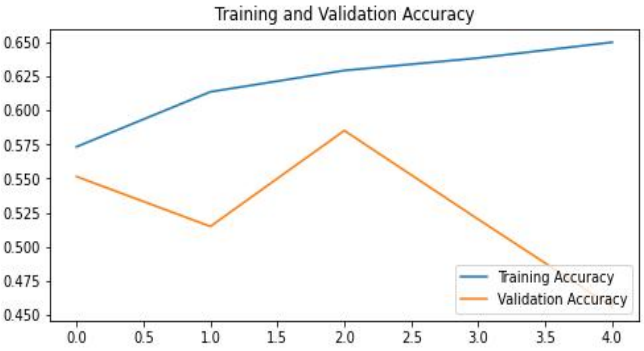
Testing

The trained Xception network was then tested with test histological images of the dataset. The overall test images are about 19230. The resulting testing accuracy is 0.566. After testing all those images are predicted correct.

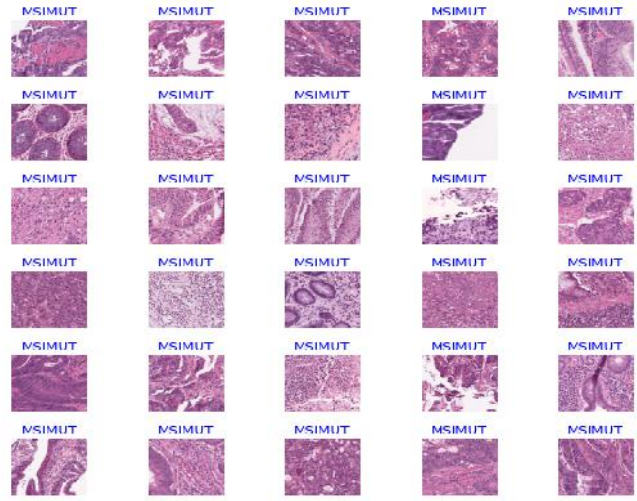
V. SCREEN SHOTS



Labeled Training Images



Training and Validation Accuracy



CNN Predictions

VI. CONCLUSION

This paper demonstrates that MSI and MSS gastrointestinal cancer classification can be most accurately done using transfer learning methodology in deep learning. A brief review of literature is performed on the classification accuracies of gastrointestinal cancer datasets and various model behaviors has been investigated. This gives an overall idea of research focus to this work and in turn facilitates to

identify the gaps and direction to undertake. The required pre-requisites have been captured and design of the deep neural network using pre-trained Xception network has been performed. The formulated pre-trained model is then trained on Google Colab research platform as training these models is compute intensive requiring GPU. The implementation of the pre-trained model illustrates the usage of GPU on Google cloud platform for training deep learning models. The algorithm has been coded in Python along with Keras API which uses the TensorFlow backend.

Future work, other gastrointestinal cancer datasets can be trained with the pre-trained Xception model developed in this work. Other pre-trained networks can also be attempted with this dataset and other datasets as well. Also, hybrid combined networks can be tried for improving AUC. The robustness of the model in this work is improved by implementing suitable data augmentation techniques which in turn enhanced the classification accuracy of the classifier. Thus this work demonstrates that MSI Vs MSSgastrointestinal cancer classification performance is reasonable to be implemented in production and more research in these areas can help patients be diagnosed accurately and especially at the right time.

REFERENCES

[1] Feng Su, Jianmin Li ‘‘Interpretable tumor differentiation grade and microsatellite instability recognition in gastric cancer using deep learning,’’ 2022.

[2] Esteva, A. et al. A guide to deep learning in healthcare. Nat. Med. 25, 24–29 (2019).

[3] Norgeot, B., Glicksberg, B. S. & Butte, A. J. A call for deep-learning healthcare. Nat.Med. 25, 14–15 (2019).

[4] Su, F. et al. Development and validation of a deep learning system for ascites cytopathology interpretation. Gastric Cancer 23, 1041–1050 (2020).

[5] Kather, J. N. et al. Predicting survival from colorectal cancer histology slides usingdeep learning: A retrospective multicenter study. PLoS Med. 16, e1002730 (2019).

[6] Hu, Y. et al. Deep learning system for lymph nodes quantification and metastatic cancer identification from whole-slide pathology images. Gastric Cancer 24,868–877 (2021).

[7] Kather, J. N. & Calderaro, J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nat. Rev. Gastroenterol. Hepatol. 17, 591–592 (2020).

[8] Arai, T. et al. Frequent microsatellite instability in papillary and solid-type, poorlydifferentiated adenocarcinomas of the stomach. Gastric Cancer 16, 505–512 (2013).

[9] Sugimura, H. Editorial: an obsession with subtyping gastric cancer. Gastric Cancer 16, 451–453 (2013). vol.Cancer Genome Atlas Research Network. Comprehensive molecular character ization of gastric adenocarcinoma. Nature 513, 202–209 (2014).

[10] Kanesaka, T. et al. Clinical predictors of histologic type of gastric cancer. Gas trointest. Endosc. 87, 1014–1022 (2017).

[11] Kuwata, T. et al. Establishment of novel gastric cancer patient-derived xenografts and cell lines: pathological comparison between primary tumor, patient-derived, and cell-line derived xenografts. Cells 8, 585 (2019).

[12] Feng, F. et al. Prognostic value of differentiation status in gastric cancer. BMC Cancer 18, 865 (2018).

[13] Liu, S. et al. Apparent diffusion coefficient value of gastric cancer by diffusion weighted imaging: Correlations with the histological differentiation and Lauren classification. Eur. J. Radiol. 83, 2122–2128 (2014).

[14] Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. 25, 1054–1056 (2019).

[15] Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. Nat. Commun. 8, 15180 (2017).