Classification of Anomalies in Gastrointestinal Tract through Endoscopic Imagery with Deep Learning

I. INTRODUCTION

The Gastrointestinal tract is an important organ in humans which digest and absorb ingested nutrients. A large number of diseases and complications can occur in the GI tract and anomaly detection is vital in diagnosing such diseases. This is even more crucial since 3 of the most common type of cancers is also related to the gastrointestinal tract.

At present, diagnosis of GI tract anomalies is done through observation of gastrointestinal endoscopic imagery by gastroenterologists. Diagnosis depends on the level of expertise of the gastroenterologist and this can lead to differing conclusions as well as misdiagnosis. Inaccurate identification of anomalies can lead to multiple complications which may be fatal to the patient and therefore accurate classification of anomalies is crucial.

It is imperative that a computerized system capable of accurately identifying GI tract anomalies be utilized for diagnosis. This will help in accurate identification of anomalies as well as being useful in areas where specialized gastroenterologists are scarcely found. Such a computerized system based on convolutional neural networks and transfer learning is proposed in this report. We initially used the KVASIR dataset to train a base model with no transfer learning and then used the same dataset to train models with transfer learning using VGG 19, Inception V3, DenseNet 201 and a Stacked Model and compared the results to obtain the best model.

II. LITERATURE REVIEW

In conducting the literature review we came across a research paper written by a group of lecturers from the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka [1] which focuses on identifying and classifying anomalies in the gastrointestinal tract through endoscopic imagery by an autonomous system with the help of deep learning. The researchers have used Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) in this research to develop the system.

The data set used in this research is KVASIR dataset. It is composed of endoscopic imagery from inside the GI tract, which has 8 different classes. It contains anatomical landmarks; z-lines, pylorus, and cecum and pathological finding; esophagitis, polyps, and ulcerative colitis, which are explained in detail in the paper. These anatomical landmarks are easily detected through the endoscope and can be used as a reference point to describe the location for findings. The dataset has 8000 images where 1000 images belong to each class. The researchers have split the dataset into 2 and fed 80% of the data as a training set and 20% as test data to the models.

The methodology adopted by the researchers is to use pre-trained CNN feature extractors which are fed into an independent classifier to get predicted class labels. The dataset is preprocessed initially and then a set of prominent CNNs with global average pooling are used to obtain feature vectors. The feature vectors are fed into a single hidden layer ANN with 128 hidden units with the ReLU activation function. SoftMax activation function has been used for the classification layer to acquire the best classification accuracy of 97.38%. Repeated stratified five-fold cross-validation is used to train the model.

The best classification accuracy was obtained by using a combination of 3 CNN feature extracts named DenseNet-201, ResNet-18, and VGG-16. They propose to use these extracts in collaboration with GAP and SVD to increase classification accuracy. The researchers have provided results to show that the proposed approach performed well with a higher classification accuracy of over 97%.

Also similar to the above paper, Samira Larfraxo and Mohamed El Ansari from the Faculty of Science of Ibn Zohr University of Morocco very recently published the paper “Abnormalities Recognition in Gastrointestinal Tract through Endoscopic Imagery using Deep Learning Techniques”. According to the paper which was published in October 2020 the researchers propose a novel deep learning model based deep convolutional neural network. The main functionality of the proposed model is to detect gastrointestinal diseases from endoscopic images. The new architecture design was validated on the KVASIR dataset which contains eight thousand images. The model achieved an accuracy of 96.89%. [2]

Andrea Asperti and Claudio Mastronardo have written a paper on the effectiveness of data augmentation [3] related to our project topic. Their approach is an ensemble of models that are created by using transfer learning from previously trained CNN nets and data augmentation. These researchers have also used the KVASIR dataset which was used in the previously mentioned research papers. (They have used the first version which has 4000 images). A transfer learning approach has been used to save computation time and to focus on the high-level representations learned by CNNs. They have used the Inception v3 model [5] and Keras library backed with TensorFlow as the backend. Their methodology was to load pre-trained weights

learned on the ImageNet dataset and after the last convolution layer to add a global averaging pooling layer, a dense layer with 1024 neurons with ReLU as activation function, and a SoftMax layer of 8 neurons. They froze all Inception's already trained layers and used Adam optimizer to tune the last dense layers' weights. They have switched to stochastic gradient descent with moment which enabled them to use a very small learning rate of 0.0001 to ensure that the magnitude of updates stayed very small and did not break previously learned features.

Apart from this, the research is significant due to the fact that the researchers have adopted several data augmentation techniques. They have used Keras' utilities to augment training instances by applying several random transformations. A list of data augmentation transformations used during the training is shown in table 1 of the paper. They have adopted a nearest pixel policy which repeats the nearest pixel value across the axis to fill pixels due to zooming out and shifting. To normalize both training and test data they have divided every pixel's color value by 255 in order to have all pixel values in the range [0,1]. This approach has enabled them to make the model more robust and prevent overfitting.

Another technique for optimizing used by them is Snapshot Ensembling. It is a method where a single model is allowed to converge into several different local minima along its optimization path on the error surface. It allows the model to dive into a minima using a decreasing learning rate, save the snapshot at that minimum and then increase the learning rate to escape the local minima in order to find other better minima. Cosine annealing cycles have been used as the learning rate schedule. They have used an initial learning rate of 0.1 and trained for 22 epochs and they have chosen an ensemble with 5 models. This Snapshot Ensembling technique helps to improve classification precision and avoid getting trapped in a local minima. This research has shown that a simple combination of CNNs, transfer learning, and data augmentation greatly improves precision and recall while maintaining the same level of accuracy. Their experimental results conclude that data augmentation can be used to boost deep learning for a small dataset.

We also came across a research paper written by 11 researchers primarily from the University of Virginia on how deep learning can be used for disease detection in gastrointestinal biopsy images [4]. The problem they address in this paper is the difficulty in distinguishing patterns when associated conditions share histological features which can lead to the wrong diagnosis. They have primarily focused on two diseases: Environmental Enteropathy (EE) and Celiac Disease (CD). The dataset used is 464 high-resolution whole slide images taken from 150 H&E duodenal biopsy slides. They have split the image slides into 1000x1000 pixel patches with an overlap of 750 pixels on both axes to ensure that all areas of the image are covered. In order to balance the dataset, they have created 4 duplicate patches by rotating one EE patch. Test Time Augmentation was done to the entire dataset to avoid overfitting the model.

The authors have mentioned that the residual learning architecture (ResNet) presented by He, Kaiming, et al. [6] that has skip connections that add outputs from a layer to the outputs from a downstream layer allows the model to be arbitrarily deep without degrading accuracy on the ImageNet dataset. They pre-trained their model on the ImageNet dataset and used transfer learning to incrementally train it on the data. A ResNet50 architecture was used to classify the patches. The learning rate used while training earlier layers is 1/9th the layer of final layers and the middle layers use 1/3rd the learning rate of final layers.

The model has reported an accuracy of 97.6% which is quite high, and the misclassification has mainly occurred in classifying Celiac and Normal images. Written by Dimitris K. Iakovidis, Spiros V. Georgakopoulos, Michael Vasilakakis, Anastasios Koulaouzidis, and Vassilis P. Plagianakos the paper "Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification" proposes a methodology that can be used for "automatic detection and localization of gastronomical anomalies in endoscopic video frame sequences".[7] In the paper researchers mainly examine weakly supervised learning for automated video analysis in Gastrointestinal Endoscopy (GIE) to deal with the resource-demanding nature of detailed image annotation. The above-discussed methodology is implemented in three main phases which are: Classification of video frames(images) into categories normal and abnormal using a Weakly Supervised Convolutional Neural Network, detecting salient points from deeper WCNN layers with the help of a novel Deep Saliency Detection (DSD) algorithm and finally localizing gastrointestinal anomalies by means of a novel Iterative Cluster Unification (ICU) algorithm. The CNN proposed to consist of five convolutional layers, by a max-pooling layer following the first four convolutional layers and by another three fully connected layers following the fifth convolutional layer. The number of convolutional layers is limited to five as it provides the best balance between the depth of the network and the overall classification performance for the specific input image dimensions. Here weakly annotated images at the image-level are used to train the WCNN. The two output neurons of the WCNN will output one of two binary vectors (0,1) or (1,0) indicating whether the image(frame) contains an anomaly or not respectively. Then the novel DSD presented in the paper is applied on both abnormal and normal training images to detect salient points. After that, the proposed novel ICU extracts Pointwise Cross-Feature-Map features and utilizes them to form clusters, named as abnormal and normal. The two datasets: MICCAI Gastroscopy Challenge Dataset and KID dataset 2(a balanced subset) were used to evaluate the proposed WCNN architecture. They were able to achieve 96% AUC (Area Under the receiver operating Characteristic) for anomaly detection and 88% AUC for anomaly localization. In summary, the methodology recommended in this paper offers a cost-efficient solution for automatic analysis GIE of video frame sequences to allow both detections, localization of gastrointestinal anomalies [7]

Another paper proposing a method for classification and identification of gastrointestinal diseases using endoscopy videos was published in 2019 by a group of researchers in the Division of Electronics and Electrical Engineering at Dongguk University in Korea. The proposed method includes a complete AI-based framework for the classification of multiple GI diseases using endoscopic videos, that has the ability to extract spatial features as well as temporal features for greater classification performance. There are four main phases of the implementation of the proposed system. They are the preprocessing phase, spatial feature extraction phase, temporal feature extraction phase, and the classification phase. In the preprocessing phase, each frame was adjusted to 224*224*3 which is according to the input layer size of the CNN (convolutional neural network) model. For the next phase, ResNet18 CNN model which consists of 5 identity mapping units and 3 convolutional mapping residual units is used for spatial feature extraction. The ResNet 18 model was selected after comparing it with other models as it gave the best learnable parameter number and the maximum accuracy. After that for the extraction of temporal features, there is Long Short-Term Memory (LSTM) based deep neural network. The feature vector obtained from the earlier step is used as the input for this step and for n spatial feature vectors n LSTM cells are used to extract temporal features. Finally, an additional neural network model is employed to perform the classification where the output of each LSTM cell traverse through a FC layer then a soft-max layer and ultimately a classification layer. The endoscopic videos from Gatrolab and KVASIR datasets containing 52,471 frames were used for experiments. The proposed model classifies the frames into thirty-seven categories using spatial as well as temporal features. In the experiment proposed model achieved 97% AUC. In summary this paper introduces a novel framework which is ResNet and LSTM-based for classification of endoscopic frames to identify different types of gastrointestinal diseases.[8]

The paper "Classification of Polyps in Capsule Endoscopic Images using CNN" published in 2018 proposes a fully trained convolutional neural network with a fully connected network for classification of endoscopic polyps. Polyp is one of the gastrointestinal abnormalities' worth identification at early stages. One of the main difficulties the team faced was the lack of labeled data and the team overcame this problem by using data augmentation and manual labeling. In the preprocessing phase, the image frames from publicly available endoscopic videos were taken and they were rotated and flipped to obtain a large dataset. All the frames were resized to 128x128x3 and normalized. All the images were manually labeled and were verified by an expert. The proposed CNN contained four convolutional layers each of the first three followed by a max-pool layer and a fully connected network with two layers following the last convolutional layer. Ultimately for the classification, the output of the last fully connected layer is connected to a SoftMax layer. ReLu function was used to deal with the vanishing gradient and Adam-optimizer was used to overcome gradient descent getting stuck at local minima. The manually labeled dataset with 31525 image frames: 15650 of bening class and 15,875 of the malignant class was used for training and validation. Here 20% of each data class was used for validation and the rest for training. Initially, the team achieved a training accuracy of 92% and validation accuracy of 87%. However, after tuning the hyperparameters and some modifications to the topology the team was able to achieve overall accuracy of 99.85% and a validation accuracy of 98.3%. In this research, transfer learning was not an option as "medical images are completely different from data coming from other domains" according to the authors.[9]

According to the previous research using Convolution Neural Networks combined with transfer learning to reduce the training time and also to improve accuracy is one of the best approaches. Furthermore, we can use stacked models to improve the accuracy of the model, the most used base models are Inception nets and ResNets. Further Augmentation of the images such as rotations, batch-wise normalizing, sample-wise normalization, shifts will improve the accuracy of the model. Since the KVASIR dataset is having an equal number of images for each class no need for under-sampling or oversampling. Therefore, in conclusion from the above research, we can conclude that using a CNN with transfer-learning will lead to a more accurate and robust model.

III. METHODOLOGY

For classifying the images of the KVASIR dataset several models were used and the results were compared to obtain the best model out of the developed models.

Initially a **base model** was developed without using transfer learning. The dataset was split into 70% training data, 15% validation data and 15% test data. Also, images were resized to 224x224 and features were normalized. The basic model consisted of 4 convolutional layers with Re-Lu activation and 4 max pooling layers placed alternatively. Dropouts of 0.2 were done after second and third convolutional layers. The vectors obtained from the above layers were flattened and passed into a 128 dense layer with Re-Lu activation. Finally, a dense layer was used for classification with a SoftMax activation function. Then 50 epochs were run using Adamax optimizer.

For the next model we used transfer learning. The next model that was developed involved using **VGG 19** features. The KVASIR dataset was split into 80% training data, 20% testing data and all the images were resized to 224x224 and features were normalized.

The pretrained model VGG19 was taken, and the last 30 layers were set as trainable. Then the output from those layers were flattened and was provided as input to a 1024 dense layer with Re-Lu activation. Then that layer was followed by a 512 dense layer with Re-Lu activation with 0.4 dropout. Finally, a dense layer was used for classification with a SoftMax

activation function. Then 50 epochs were run using SGD optimizer.

For the next transfer learning model, we used Inception V3. The KVASIR dataset was split into 80% training data, 20% testing data and all the images were resized to 224x224 and features were normalized.

The pretrained model **Inception V3** was taken, and the last 30 layers were set as trainable. Then the output from those layers were flattened and passed into a 1024 dense layer with Re-Lu activation. Which was then followed by a 512 dense layer with Re-Lu activation with 0.4 dropout. Finally, a dense layer was used for classification with a SoftMax activation function. Then 50 epochs were run using SGD optimizer.

The next model involved feature learning using **DenseNet201**. The KVASIR dataset was split into 70% training data, 15% validation data, 15% testing data and we use the keras Datagenerator to load the data into memory during the three phases. We add small data augmentations like shifts and rotations.

We are using the pretrained (using imagenet dataset) Densenet201 initially and by disconnecting the convolution block 5 and attaching two 512 feature convo layers and 3 dense layers for classification (using dropout layers in the middle) we initially train the newly attached layers while keeping the original layers non trainable to train the newly added layers a bit and then in the second round train the last 150 layers all together.

We also developed a **Stacked Model** where we are using our custom and Inceptionv3 models and getting the feature vectors/tensor at one layer before the prediction layer. From this we get 2 (,128) sized tensors. Using a concat layer we are combining the 2 feature vectors (,256) and sending it through 3 dense layers to make an ensemble from the two models.

IV. CLASSIFICATION EXPERIMENTS AND DISCUSSION OF RESULTS

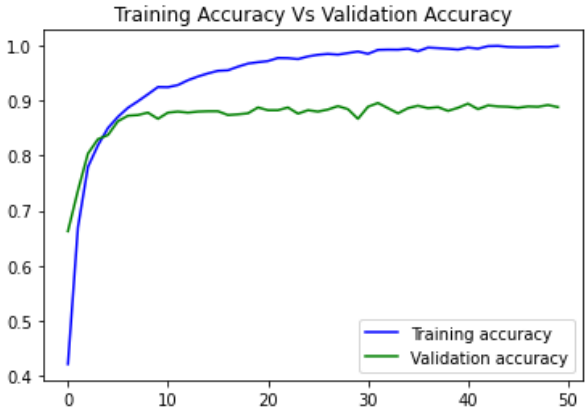
Initially a base model with no transfer learning was developed to establish a baseline for the accuracy followed by 4 other models with transfer learning using VGG 19, Inception V3, DenseNet201 and a Stacked model. The training and validation accuracies obtained from the above-mentioned models are tabulated as follows.

	Training Accuracy	Validation Accuracy
Base Model	65.31%	55.34%
VGG 19	99.87%	88.75%
Inception V3	99.75%	87.88%
DenseNet 201	99.30%	80.50%
Stacked Model	98.07%	89.08%

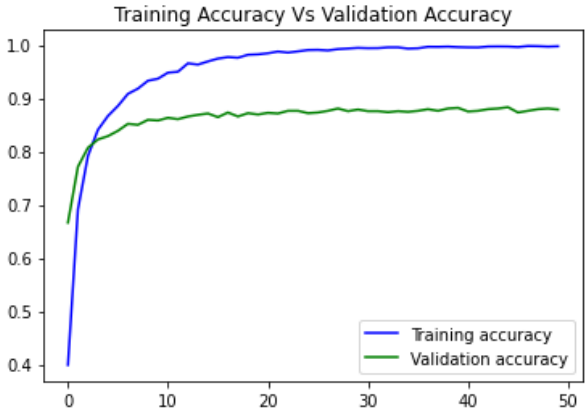
The test accuracy from the Stacked model was 90.00%.

The accuracy graphs obtained from the models are given below.

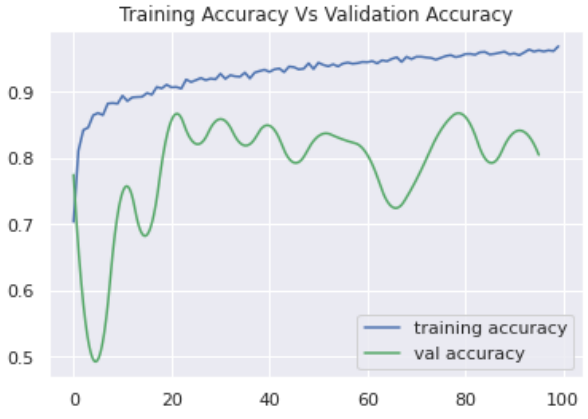
1. VGG 19



2. Inception V3

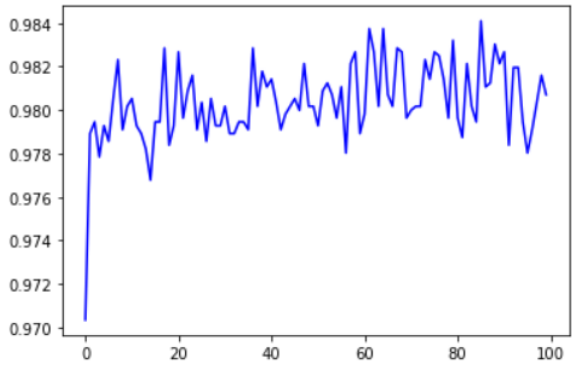


3. DenseNet 201

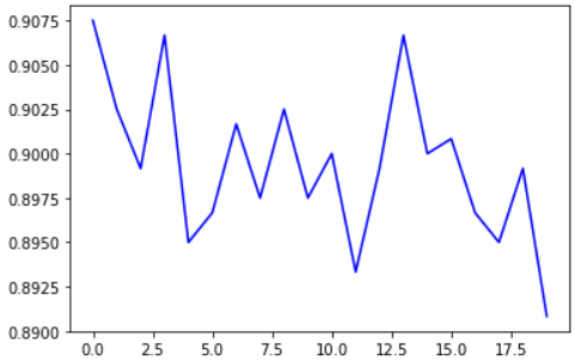


4. Stacked Model

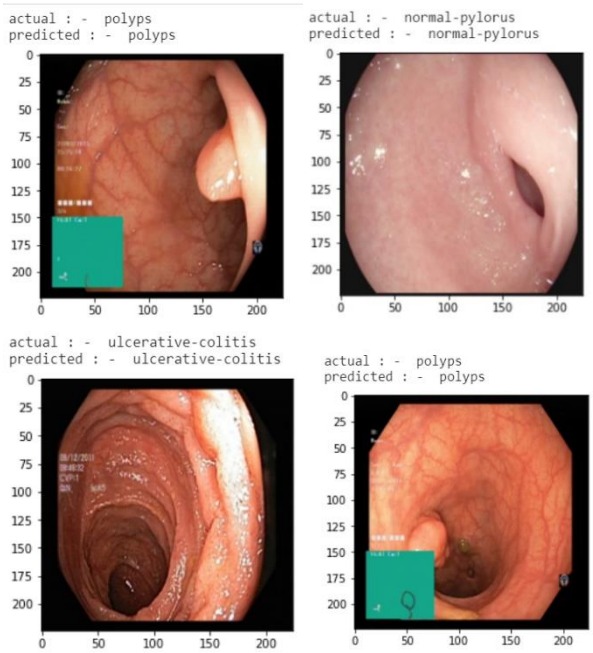
Training accuracy



Validation accuracy



Classifications made by the stacked model.



V. CONCLUSIONS

We have developed a base model with no transfer learning and 4 models with transfer learning using VGG19, Inception V3, DenseNet201 and a Stacked model and compared the results obtained when classifying the images of the KVASIR dataset. While the model developed using VGG19 was a close second with 88.75%, the best validation accuracy of 89.08% was obtained from the Stacked model. A test accuracy of 90.00% was obtained from the Stacked model which is an ensemble of our custom model and the InceptionV3 model.

REFERENCES

[1] C. Gamage, I. Wijesinghe, C. Chitraranjan and I. Perera, "GI-Net: Anomalies Classification in Gastrointestinal Tract through Endoscopic Imagery with Deep Learning," 2019 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2019, pp. 66-71, doi: 10.1109/MERCon.2019.8818929. [online] Available at: https://ieeexplore.ieee.org/document/8818929

[2] S. Lafraxo and M. El Ansari, "GastroNet: Abnormalities Recognition in Gastrointestinal Tract through Endoscopic Imagery using Deep Learning Techniques," 2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM), Reims, 2020, pp. 1-5, doi: 10.1109/WINCOM50532.2020.9272456[online]Available at https://ieeexplore.ieee.org/document/9272456

[3] A. Asperti. and C. Mastronardo, “The Effectiveness of Data Augmentation for Detection of Gastrointestinal Diseases from Endoscopical Images.” 2019 [online] Available at: https://arxiv.org/abs/1712.03689

[4] A. Srivastava *et al.*, "Deep Learning for Detecting Diseases in Gastrointestinal Biopsy Images," *2019 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, 2019, pp. 1-4, doi: 10.1109/SIEDS.2019.8735619 [online] Available at: https://ieeexplore.ieee.org/document/8735619

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking The Inception Architecture For Computer Vision”. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2818– 2826. [online]Available at: https://ieeexplore.ieee.org/document/7780677

[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.[online]Available at: https://ieeexplore.ieee.org/document/7780459

[7] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis and V. P. Plagianakos, "Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification," in IEEE Transactions on Medical Imaging, vol. 37, no. 10, pp. 2196-2210, Oct. 2018, doi:

10.1109/TMI.2018.2837002. [online]. Available at https://ieeexplore.ieee.org/document/8359121
[8] O. Muhammad, A. Muhamaad, J. Choi, T. Mahmood and K. Park (2019). “ArtificialIntelligence-Based Classification of Multiple Gastrointestinal Diseases Using Endoscopy Videos for Clinical Diagnosis”. Journal of Clinical Medicine. [online] Available at: https://www.mdpi.com/2077-0383/8/7/986/html
[9] P. Sasmal, Y. Iwahori, M. K. Bhuyan and K. Kasugai, "Classification of Polyps in Capsule Endoscopic Images using CNN," 2018 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, India, 2018, pp. 253-256, doi: 10.1109/ASPCON.2018.8748527. [online] Available at: https://ieeexplore.ieee.org/document/8748527